Heterogeneity and Decision-Making in Cellular Signaling Networks

By

Ryan Suderman

Submitted to the Center for Computational Biology and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Committee members

Eric J. Deeds, Ph.D., Chairperson

J. Christian Ray, Ph.D.

John Karanicolas, Ph.D.

Wonpil Im, Ph.D.

Chris Fischer, Ph.D.

Date defended: March 30, 2016

The Dissertation Committee for Ryan Suderman certifies
that this is the approved version of the following dissertation :

Heterogeneity and Decision-Making in Cellular Signaling Networks

Eric J. Deeds, Ph.D., Chairperson

Date approved: March 30, 2016

Abstract

Signaling networks are the means by which cells adapt to their environment. Over the last decade, cellular decision-making has been shown to exist in the midst of substantial heterogeneity, even among isogenic cells. The presence of such variability is generally assumed to be an obstacle that cells overcome in order to precisely resolve information about their environment. In this work, we investigated the effects of two specific types of intracellular heterogeneity on signal transduction in cells. The first is the presence of substantial compositional heterogeneity in the sets of macromolecular complexes used for signal transduction, which we observe in a model of the yeast pheromone signaling system. In spite of this, the model is able to reliably reproduce experimentally observed dynamical and dose-response trends. We then contrasted this model with one that employs a hierarchically assembled, stable signaling complex and found that the two signaling paradigms can exhibit distinctive behaviors. These differences can be attributed in part to the role of the scaffold protein in signal complex assembly, which is required for signal transduction in the pheromone network. We found that features such as signal amplification and crosstalk prevention vary depending on how the assembly of scaffold-based signaling species occurs. Our results clearly show that a dynamical understanding of signal transduction must take place in the context of compositional heterogeneity. The second form of heterogeneity we consider, biochemical noise, occurs at a more fundamental level. We examined how variability in the response to signal impacts the ability of cells to make reliable decisions by quantifying signal transduction using concepts from information theory. Our results revealed the existence of a fundamental trade-off: increased noise in individual

cells corresponds to increased information available to control cellular populations. To provide context for the general application of information theory to cell signaling, we characterized the upper limits of information transmission in models of simple signaling motifs. Our results also revealed that certain features of signaling networks, such as enzyme saturation and molecular copy number, are central to regulation of information transmission through networks of arbitrary size. With formal, systematic modeling approaches, we were able to elucidate many non-intuitive behaviors resulting from variability in signal transduction. Thus, we expect that our treatment of heterogeneity in signaling networks will form the basis for the development of a comprehensive theory of cellular decision-making.

Acknowledgements

A number of individuals have been instrumental in my completion of this dissertation, and I would like to acknowledge them here.

The most notable is of course my advisor, Dr. Eric Deeds, who guided me through my career as a graduate student at the University of Kansas. His limitless supply of ideas drove my projects, but did not infringe on my own desires for what I wanted to accomplish. Because of him, I was introduced to the q-bio conference and the associated community of scientists, which spurred the development of my own scientific career. He was essential in my development as a writer and taught me how to express scientific ideas both thoroughly and concisely. Most importantly, though, is his extremely intense excitement for science, which was necessary to keep me going on occasions when my own excitement waned.

I would like to thank the individuals serving on my defense committee, Dr. Wonpil Im, Dr. J. Christian Ray, Dr. John Karanicolas, and Dr. Chris Fischer, and the remaining faculty members in the Center for Computational Biology, Dr. Joanna Slusky and Dr. Ilya Vakser, for their critiques of my work. Debbie Douglass-Metsker provided excellent administrative support both throughout graduate school and in facilitating the paperwork necessary for completion of this dissertation. The current and former members of the Deeds lab, including Maulik Nariya, Addison Schauer, Koan Briggs, Anupama Kante, Zaikun Xu, Dustin Maurer, and Dr. Michael Rowland also provided excellent feedback on my research throughout the years.

I had excellent professors as an undergraduate who were central to my success as a graduate student. I would like to thank my undergraduate advisor, now retired from the Goshen College Mathematics department, Dr. Ron Milne. Perhaps the most major contribution to my trajectory towards graduate school came from Dr. David Housman, who took the time to engage with me in an independent study of mathematics in biology.

My family has always been supportive of my choice to pursue higher education, so I would like to thank my parents Mark and Elaine Suderman for all that they have done to assist me throughout graduate school.

Finally, I could not have come this far without my wonderful wife Erin. She has sacrificed so much time and effort to support me in my studies, whether taking care of our son, Luke, consoling me during setbacks, or celebrating successes. Everything I have achieved during graduate school is because of her.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Signaling networks are the means by which cells communicate with their environment. Traditionally, these networks of interacting macromolecules, primarily composed of proteins, have been divided up into "pathways" with some well-defined input, output and set of intermediate signaling species. These signaling species often represent multi-subunit protein complexes that can potentially exist in a number of chemical states. Representations of these pathways and complexes in scientific literature generally appear in the form of a large, multi-subunit macromolecular complex (much like a ribosome, but for signal transduction) and the described output of the pathway is usually construed as a well-defined behavioral change by the cell ("activation of transcription factor X") [1]. These depictions have, both implicitly and explicitly, promoted the perspective that reliability and robustness in signal transduction is equivalent to orderly activity within these networks, with a general suppression of biochemical noise and discrete, structurally well-defined signaling complexes [2, 3].

Over the last couple of decades, however, signs have pointed toward the existence of significant non-genetic heterogeneity in cells in two distinct, but related features of signaling networks. One of these features, biochemical noise, has been well documented since the early 2000's [4, 5]. This term generally refers to the presence of variability in a molecular observable to some distinct extracellular stimulus in a population of isogenic cells. The presence of both intrinsic and extrinsic

sources of noise can give rise to fluctuations in signaling proteins and possibly wide distributions of molecular responses [6–9]. Another is the potential for structural and compositional heterogeneity in the set of signaling species employed for signal transduction [1, 10–13]. This derives in part from the astronomical number of potential molecular species that can be generated by the network, since its constituent macromolecules can exist in many possible binding or modification states. With these two types of extreme uncertainty in signaling networks, it is unclear how cells reliably make decisions affecting their own fate. Through novel theoretical and computational approaches, the work described in this thesis has begun to address the issue of reliable signal transduction with regard to these two aspects of non-genetic heterogeneity.

The first half of this work (Chapters 2 and 3) focuses on heterogeneity in protein complex formation as a result of combinatorial complexity, and the effects that various assembly paradigms (*i.e.* the mechanisms by which multi-subunit complexes are formed) have on key properties of signaling networks. This feature of signaling networks and the resulting consequences for the networks' dynamics have been identified for a number of years [10], and as a result, a number of hypotheses as to the nature of complex formation have developed [11]. However, until recently, the computational tools for simulating and analyzing a model of such a complex signaling network had yet to be developed. The advent of rule-based modeling [14, 15], coupled with methods for stochastically sampling the space of all possible chemical species, based on the Doob-Gillespie algorithm [16–18] made this computationally possible (Figure 1.1).

We thus explored the feasibility of signaling via *pleiomorphic ensembles*, first proposed by Mayer, *et al.* [11], which consist of diverse sets of transient molecular species and do not require any sort of structural or compositional order. In essence, this means that there is not some "core" signaling complex responsible for transmission of information; local interactions that are independent of other molecular context (*e.g.* binding of *A*'s SH2 domain to *B*, regardless of *A*'s other domains, Figure 1.1) are sufficient for signal transduction. It is this independence that gives rise to combinatorial complexity in the protein interaction network, which leads to compositional heterogeneity among the sets of signaling species employed for signal transduction in isogenic

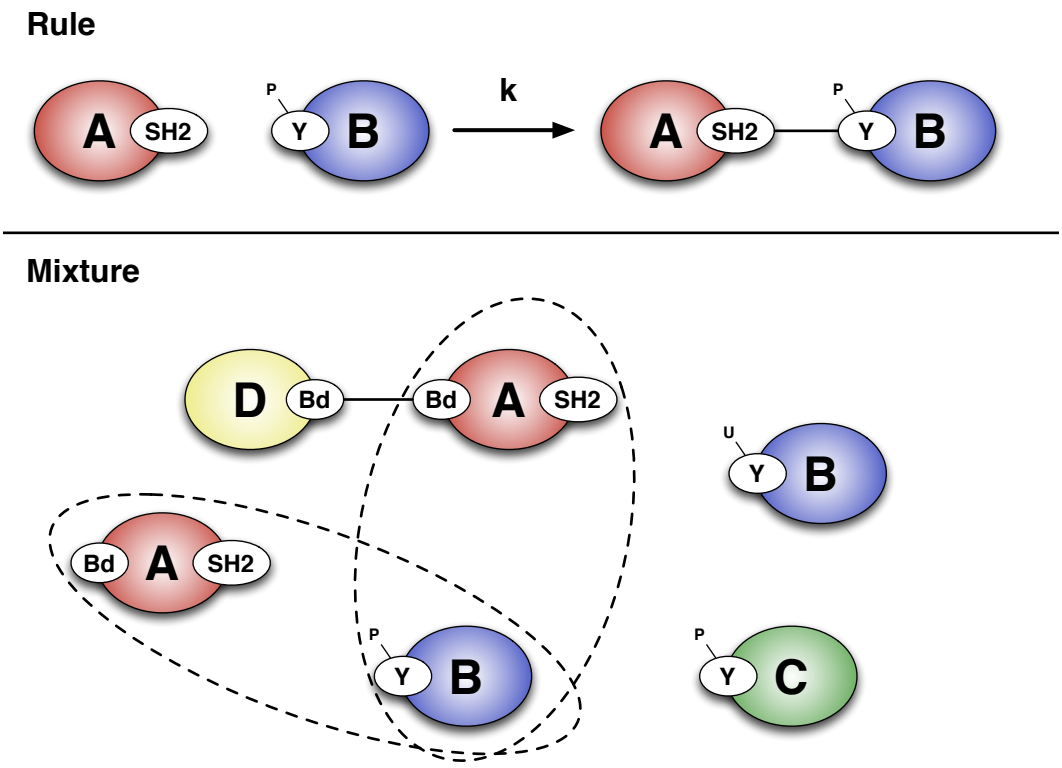


Figure 1.1: The upper panel contains an example rule describing an association event with a rate, k. Here, a site representing an SH2 domain as part of some agent (*e.g.* protein type) labeled A to a site representing a phosphorylated tyrosine (Y $\sim$ P) on some agent B. The lower panel is a *mixture* (*i.e.* a multiset of chemical species present at a particular point in time in a simulation) to which the rule can be applied if the left hand side of the rule matches patterns in the mixture. There are two possible applications, denoted by dashed lines. Note that both A and B can have other sites and states (such as the binding site, Bd, on agent A) representing certain states the agents can have that are not present in the rule; the absence of such information means that the association event occurs independently of the omitted sites and states. This allows a succinct description of a reaction network without the need for *a priori* enumeration of all chemical species as in a system of differential equations

cells. We developed a model of the yeast pheromone signaling network, whose core components include a G-protein cycle and a scaffold-dependent kinase cascade, built using only knowledge of the relevant protein interactions and assuming that all biochemical events can occur independently unless a conditional dependence was experimentally demonstrated in the literature [1]. Our results showed that the set of signaling complexes generated between independent simulations that originated from the same microstate were distinctly different on average, while producing dynamical and dose-response trends that replicated experimental data. Furthermore, this model did not produce anything resembling a "core" signaling complex, or a sub-complex that was conserved over a majority of sampled complexes. We therefore labeled this model an *ensemble-like* model and proceeded to create a contrasting *machine-like* model.

To construct a machine-like structure while maintaining all existing interactions, we introduced a hierarachical binding procedure: kinases were required to bind to the scaffold protein in a specific order (eliminating the majority of the combinatorial complexity), and we stabilized the final machine-like structure by modifying the kinetic rate constants. The machine model was also capable of replicating experimental trends, and so we examined whether or not one could possibly distinguish between the two models using experimental techniques. Ultimately we found that traditional experimental methods of characterizing protein "complexes," such as tandem affinity purification coupled with mass spectrometry, were incapable of distinguishing the two models based on the structure of the complexes they generated [19]. We demonstrated, however, that another feature of the network, *combinatorial inhibition*, provided indirect evidence of ensemble-like behavior in this system [20, 21]. This particular distinction arose, since the presence of combinatorial inhibition requires a multivalent scaffold protein to which its constituents bind independently [20], which is clearly present in the ensemble model, but not in the machine model due to its hierarchical assembly process.

A number of other functions have been proposed for multivalent scaffold proteins and a few have been investigated, but never in the context of machine- and ensemble-like complex assembly [22, 23]. We therefore proceeded to formally and systematically develop a theoretical under-

standing of how scaffold proteins impact various dose-response and dynamical features of signaling networks (see Chapter 3). One prominent example is the supposition that the stoichiometric constraints of scaffolds binding their effectors would completely suppress signal amplification [22, 23]. In fact, we found the exact opposite: assembly of machine-like signaling complexes using a scaffold as a nucleation point induces higher amplification than a kinase cascade with no scaffold. We also observe a lesser degree of amplification in ensemble models. Another example is the idea that scaffolds could prevent signal leakage (a form of *crosstalk*) between signaling pathways that share components [24]. Ensemble-like networks do mitigate some, but not all, inappropriate cross-pathway activation. On the other hand, machine-like networks completely eliminate cross-pathway activation, but one pathway can still affect another when the cascades' shared components are in limiting concentrations. Our results clearly show how seemingly intuitive reasoning about nonlinear dynamical systems can fail, and how systematic modeling approaches can contribute to our understanding of complex signaling networks.

Our results also revealed the possibility that compositional heterogeneity can be a beneficial feature in terms of evolutionary fitness. In particular, we found that machine- and ensemble-like protein interaction networks (not just those responsible for communication of extracellular signals) could have contrasting evolutionary roles. Machines may provide functional stability and robustness for core cellular processes (*e.g.* ribosome formation and translation or proteasome assembly and protein degradation). Ensemble-like networks, on the other hand, could represent a form weak linkage, exhibiting increased functional or phenotypic plasticity and promoting variation in the evolution of interaction networks [1, 25].

In contrast, the presence of biochemical noise in signaling networks appears (at least on the surface) to have minimal benefit to the cell. The second half of this work (Chapters 4 and 5) focuses on how cells can reliably interpret information about their environment, and what happens if they cannot. In particular, noise can prevent cells from having sufficient information to distinguish between two distinct levels of signal, a sort of cellular "confusion" (Figure 1.2). In order to quantify this noise and its effects on signal transduction, we applied concepts from information

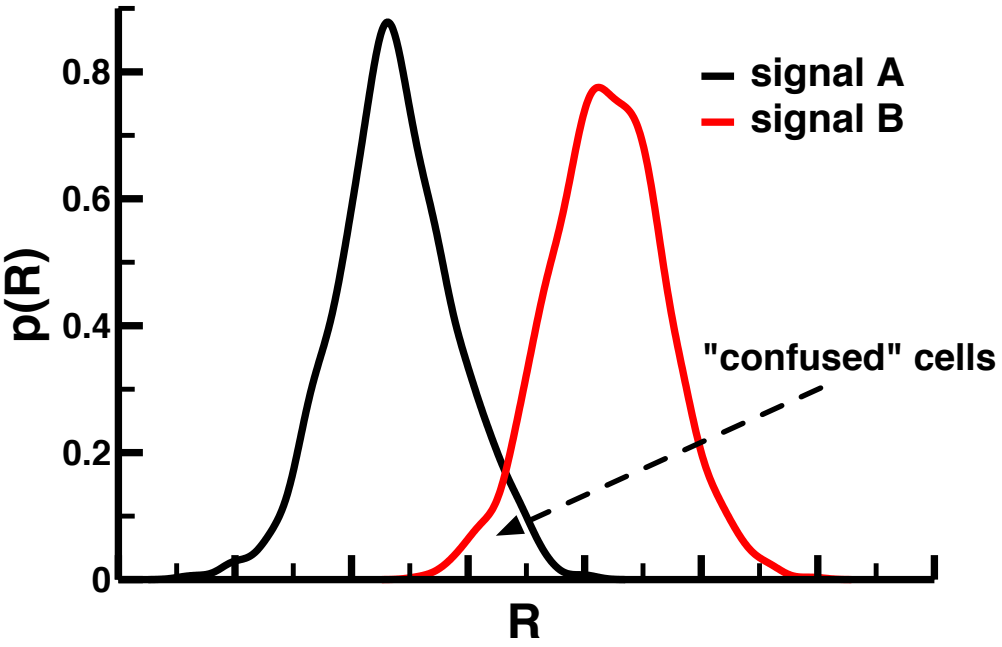


Figure 1.2: In this example, two distinct signal values produce response, R, distributions in some population of isogenic cells with sufficient variability such that the distributions overlap with each other. The cells that exist in this overlapping region are incapable of determining to which signal they were exposed and can then be considered "confused". This confusion, or level of reliability in information transmission, can be quantified using concepts from information theory.

theory, a mathematical framework designed to formally describe the communication of information over a channel that is susceptible to noise [26]. A prominent application of information theory to metazoan systems biology and signal transduction examined the *channel capacity*, or maximum amount of information that can be sent through a channel [27], of the TNF-α induced signaling network and found that the network is incapable of distinguishing between 2 distinct levels of signal [9]. In other words, the network transmitted less than 1 bit of information and is thus unable to make a binary decision. This observation has led to a number of studies examining how cells could implement molecular mechanisms to mitigate the noise responsible for this seemingly low level of information transmission [2, 28].

We hypothesized that there might be a benefit to the presence of noise in individual cells.

Using a combination of simple models and experimental investigation of the extrinsic apoptosis signaling network, we demonstrated the existence of an inherent trade-off between the amount of information transferred in individual cells and the information available to control population-level responses (see Chapter 4). In particular, reduction of information transfer at the single-cell level can *increase* the information transmission to a cellular population. Furthermore, we found that the previous low levels of information transmission characterized by Cheong *et al.* were not indicative of an upper bound on information transmission in individual eukaryotic cells [9]. For processes such as eukaryotic chemotaxis or yeast mating, in which individual cells must make precise cell-fate decisions, we find notably higher levels of information transmission (> 2 bits) at a single-cell level, sufficient for distinguishing between over 4 distinct levels of signal. We therefore propose that signaling networks can exploit high noise (or even generate it) in individual cells to maximize population-level information transfer.

The consideration of certain estimated levels of information transmission in single cells as "low" leads to the realization that, to date, there is minimal context for what constitutes high or low information transmission in cellular signaling networks. Furthermore, comparison of existing channel capacity calculations are difficult since estimating the value depends on numerous factors, such as the range and number of signal values that are sampled to characterize the underlying signal-response relationship, that have not systematically been studied. With this in mind, Chapter 5 describes our development of a theoretical understanding of signal transduction and decision-making in the context of information theoretic quantities. The first step in this process involved characterizing the intrinsic upper limits of information transmission in various simple signaling motifs. To maintain consistency between estimates in distinct signaling networks, we built a framework for reliable comparison of the estimated information transmission between arbitrary signaling systems, provided they have some realized steady-state response to signal. Analysis of the bounds of information transmission in signaling networks revealed a number of interesting features both general and specific to certain network motifs, and will likely provide intuition for analysis of larger and more complex signaling systems such as the pheromone network in yeast.

These analyses of non-genetic heterogeneity in signaling networks have shown that common intuition of what constitutes evolutionary fitness in signal transduction can be misleading. Through the application of novel computational methods and statistical tools for examining cellular heterogeneity, we were able to characterize the effects of two types of non-genetic variability on signal transduction, without undue simplification. These results will influence future work by serving as an example of quantitative and systematic analysis of complex biological phenomena, both computational and experimental, ultimately forming the basis for a theoretical understanding of decision-making and information transmission in cells.

Chapter 2

Machines vs. Ensembles: Effective MAPK Signaling through Heterogeneous Sets of Protein Complexes

2.1 Introduction

Much of our reasoning about the function of biological systems relies on the formation of multi-subunit protein complexes [3]. In some cases, such as the ribosome and the proteasome, these complexes take the form of intricate molecular machines with well-defined quaternary structures [29–31]. The overall structure of complexes formed during signal transduction, however, is considerably less clear. There are a few well-characterized signaling machines, like the apoptosome, and some have argued that the majority of structures produced by signaling networks would have a machine-like character [32, 33]. Most of the complexes formed during signal transmission and processing have not had their global three-dimensional structures experimentally determined, however, and as such we currently do not know the extent to which signaling occurs via machines [11]. Despite this uncertainty, the machine-like perspective on signaling complexes is pervasive in the literature, if often implicit; for instance, one commonly represents signaling networks graph-

ically by drawing large complexes in which all of the relevant proteins interact simultaneously [21, 22, 34–38] (Fig. 2.1A). Although such diagrams are often presented as compact summaries of a set of interactions, they are certainly evocative of a machine-like structure, and lead naturally to analogies between signaling complexes and highly ordered objects such as circuit boards [11, 22].

One issue that complicates this machine-based picture is the fact that the protein interaction networks that underlie cellular signaling exhibit considerable combinatorial complexity; that is, they can (theoretically) generate anywhere from millions to 10^{20} or more unique molecular species [10–12, 39]. For example, even a single PDGF receptor dimer has 105 possible phosphorylation states, many of which could be (stably) occupied by any given molecule [11, 40]. A similar problem arises in protein folding: a polypeptide chain could theoretically adopt so many conformations that it is *a priori* difficult to understand how a protein folds quickly and stably into a single native structure [12, 41, 42]. Proteins have evolved energy landscapes with specific features in order to overcome this problem (which is known as the "Levinthal paradox"). In order to assemble well-defined signaling machines, signaling networks would similarly need to evolve specific "chemical potential landscapes" in order to drive the system to a specific set of quaternary structures [12, 41].

Mayer *et al*. have speculated, however, that signaling networks might not need to assemble machine-like structures at all in order to function [11]. This "pleiomorphic ensemble" hypothesis posits that heterogeneous mixtures of complexes drive cellular responses to external signals. Early work, based on systems of Ordinary Differential Equations (ODEs) that considered a few hundred molecular species, indicated that more diffuse "network" models of signaling could generate reasonable signaling behavior [43, 44]. The dearth of computational methods that can handle combinatorially complex networks has made it difficult to fully test the ensemble hypothesis in realistic networks, however [36]. As such, it is currently unclear if ensembles could even produce reliable responses to signals, or if there is any functional or evolutionary difference between networks that employ ensembles vs. machines.

Over the past 10 years, a set of rule-based methods have been developed that allow one to model the behavior of biological systems without an *a priori* reduction in the set of possible species that

can be formed [12, 16, 18, 36, 43]. Given a model consisting of a specific set of protein interaction rules, we can exactly sample sets of protein complexes (or "conformations") from the astronomically large set of all possible complexes the model can generate. In this work we employed these methods to investigate the possibility of signaling via ensembles *in silico*. We focused on the pheromone response network (Fig. 2.1A), one of multiple mitogen-activated protein kinase (MAPK) cascades in Saccharomyces cerevisiae. This thoroughly characterized signaling cascade involves the scaffold protein Ste5, which is thought to be a nucleation point for the formation of signaling complexes (Fig. 2.1B) and prevent crosstalk [22, 34, 35]. Since similar MAPK cascades are found in eukaryotic cells from yeast to humans [45], this network represents an excellent model system for exploring the influence of combinatorial complexity on signaling dynamics.

In our initial model, we included only those interactions (and their requisite molecular contexts) that have been explicitly characterized experimentally. We found that this model is able to fit available data on the response of the network to pheromone, despite exhibiting significant ensemble character. We also constructed an alternative set of rules that could assemble a scaffold-based signaling machine, similar to those typically drawn to graphically summarize the cascade [21, 22, 34–38] (Fig. 2.1A). Although this model does fit some of the available data, we found that it could not replicate the "combinatorial inhibition" of the pathway observed at high levels of Ste5 overexpression [20, 21]; instead, it displayed considerable robustness to such changes. We also demonstrated that TAP/MS, a common technique for experimentally determining the components of "molecular machines" via binary interactions [19, 46], could not distinguish between the complexes formed in these two models, despite their radically different character. Direct experimental tests of the ensemble hypothesis thus require the application of assays that can measure three-way or higher-order interactions, such as fragment complementation, fluorescence triple correlation spectroscopy or single-molecule approaches [47–53]. Our findings indicate that ensembles can indeed reliably transmit and process extracellular information, and their inherent plasticity in response to perturbations like scaffold overexpression implies that they may play a role in facilitating the evolutionary variation of signaling systems within cells [25].

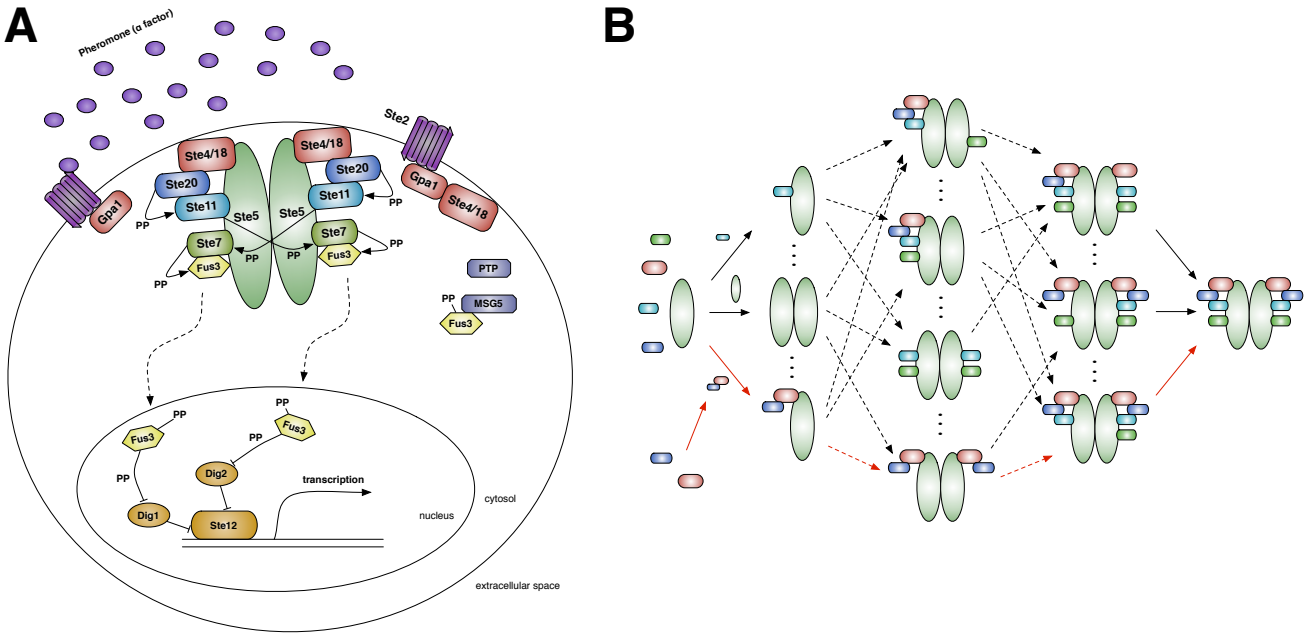


Figure 2.1: The yeast pheromone MAPK network. (A) A typical representation of the cascade. Pheromone (α-factor) stimulates G-protein activation via a GPCR (purple and red). The subsequent recruitment of the scaffold to the membrane enables the kinase phosphorylation cascade (blue and green), ultimately activating the MAPK, Fus3 (yellow), and regulating mating-related genes (orange). (B) Scaffold-based species potentially generated during our model's phosphorylation cascade (color coded to Fig. 2.1A). Solid arrows represent association events between either two monomers or a monomer and oligomer. Dashed arrows indicate a series of these association events. Red arrows indicate possible assembly pathways for the decamer (far right) in the machine model. Note that this is a very small sample of the entire set of scaffold-based signaling species and their possible interactions.

2.2 Results

2.2.1 Constructing a model of pheromone signaling in yeast

A summary of the molecular interactions underlying the yeast pheromone response network may be found in Fig. 2.1A. Briefly, the signaling cascade is initiated by the interaction between extracellular pheromone molecules and a G-protein coupled receptor (GPCR), which induces dissociation between the α subunit (Gpa1) and $\beta\gamma$ subunits (the Ste4-Ste18 complex, hereafter referred to as Ste4) of the G-protein [54]. Ste4 then recruits the scaffold protein, Ste5, which dimerizes, binds numerous kinases (Ste20, Ste11, Ste7) and promotes a phosphorylation cascade resulting in dual-phosphorylation and activation of the MAPK, Fus3 [55, 56]. As mentioned above, the vast majority of graphical depictions of this cascade involve simultaneous binding of all requisite proteins to Ste5 (Fig. 2.1A) [21, 22, 34–38], however to our knowledge there is no explicit experimental evidence that such a large scaffold-based complex is actually formed during signaling. Active Fus3 then translocates to the nucleus, regulating the expression of numerous mating-related genes via the transcription factor Ste12 [35].

To create a dynamical model of this cascade, we constructed a set of rules for these interactions and other events (*e.g.* post-translational modification, protein synthesis and degradation, nucleotide transfer). The rules themselves, which follow mass-action kinetics, were primarily derived from two sources: an online model (`http://yeastpheromonemodel.org`) [57] and an ODE model [36], both of which are based on comprehensive literature searches (Section A.1). In our initial model, if a reaction (*e.g.* efficient phosphorylation of Fus3 by Ste7) requires conditions that have been experimentally characterized (*e.g.* Ste7 also bound to Ste5), they are explicitly represented in the rule. We added no additional constraints to this model, in order to: (a) see if existing knowledge of these interactions is sufficient to produce realistic network dynamics (Fig. 2.2) and (b) characterize whether they result in machine- or ensemble-like character. The rule set, written in the Kappa rule-based modeling language [58], contains 232 rules, 18 protein and 8 gene agent types. This model displays considerable combinatorial complexity: even if we only focus on

complexes containing the Ste5 scaffold, the system can generate over 3 billion unique molecular structures (Section A.3.5). We thus employed KaSim, an open source simulator for Kappa models, to consider the dynamics of the system without a reduction in its combinatorial complexity. Our general simulation strategy is described in detail in Section 2.4.1 and Section A.2; a graphical schematic can be seen in Fig. 2.3A

2.2.2 Parameterization of the model

The model described above has two types of parameters: initial copy numbers (*i.e.* concentrations) for each of the 18 protein agents and stochastic rate constants for each of the 232 rules. We obtained the initial conditions directly from experimental measurements of copy number in yeast cells [57, 59]. The stochastic rate constants were obtained from a combination of experimental data and parameter fitting. Briefly, 7% of the rate constant parameters in the model have been directly measured for yeast proteins, 68% were estimated from measurements on related proteins in other networks and 25% were completely unknown and thus given approximate values. In order to reproduce experimental observations with our model, we identified 111 rules that were likely to influence experimentally characterized trends and varied their rate constants.

We found that only 25 of these parameters had a strong impact on the dynamics of important observables in the model, and so we modified only those values during our fitting procedure. Of these 25, 22 had original estimates obtained from related proteins. In those cases, we restricted variation of the parameters to an increase or decrease of about one order of magnitude, to maintain similarity between the fitted value and the original estimate. Two of the remaining parameters had no available estimate, and so we restricted variations in those parameters to a biologically realistic range (a table with ranges for each type of parameter is available in Section A.1.2). Finally, one parameter, the Gpa1 degradation rate, had been measured experimentally; we restricted variation in this parameter to a less than five fold change, a reasonable range given the inherent error in the experimental measurement [60]. Further details on how we identified and varied these parameters may be found in Section A.1.2.

Since each simulation of this model requires over three hours of CPU time, we could not perform fits using standard techniques, nor could we employ statistical methods to understand the probabilistic structure of the parameter space [61, 62]. Therefore, we manually altered these 25 parameters (subject to the above constraints) and simulated the model with the updated rate parameters. We iteratively applied this procedure until the model successfully replicated the dose-response behavior of Fus3 with respect to pheromone (Fig. 2.2A) [37, 38], the temporal dynamics of G-protein activation (Fig. 2.2B) [54], and other experimental observations (Figs. A.2, A.4, and A.5). To test the robustness of our results to the particular simulation method, we translated our rules into the related BioNetGen Language (BNGL) and used the same parameters to simulate the model using the BNGL simulator NFsim [18]. The two software packages produced exactly the same dynamics for these rules (Figs. A.1, A.2, A.4, and A.5 and Section A.2.2).

Given the large number of parameters in the model compared to the amount of data available for fitting, one should not construe the above results as implying this model represents a uniquely valid description of the system. Indeed, as we demonstrate below, even fairly different rule sets can provide (roughly) equivalent fits to this data; we thus cannot make any claims regarding the identifiability of the parameters or even the rule set itself [62, 63]. The point in this case is that it is possible to find some set of parameters that replicate the data, indicating that this model is at least consistent with available observations.

2.2.3 Heterogeneity in signaling complexes

To determine if the model described above signals through ensembles, we implemented a pairwise comparison between the sets of complexes produced in two independent simulations i and j, using the Jaccard distance, which we refer to as "compositional drift" [12]:

$$d(i,j) = \frac{|C_i \triangle C_j|}{|C_i \cup C_j|} \tag{2.1}$$

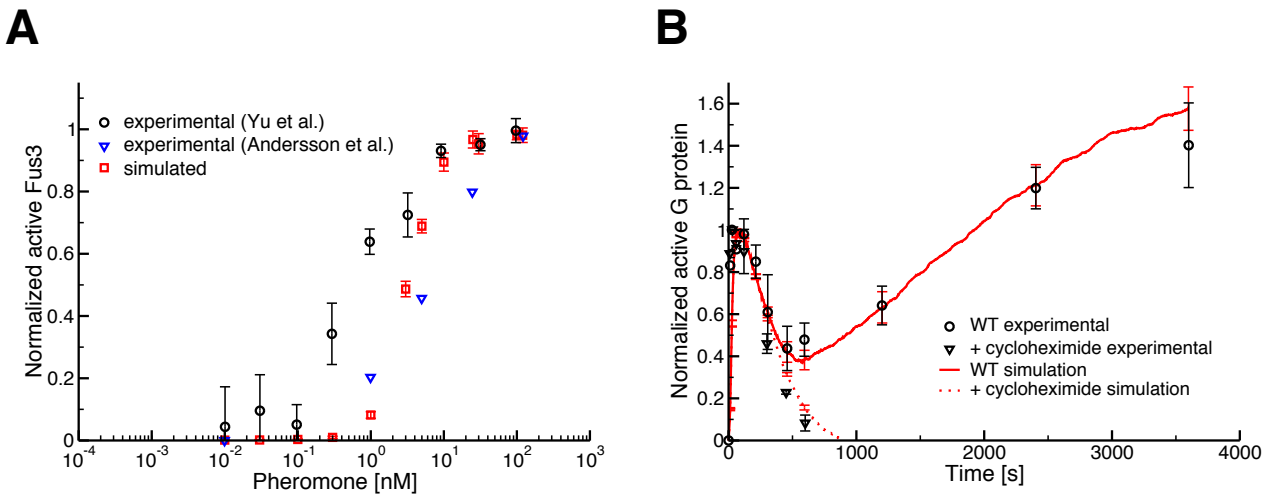


Figure 2.2: Experimental validation of the rule-based ensemble model. All error bars are 95% confidence intervals (simulated $n = 10$ where the simulations start from identical steady-state initial conditions). (A) Dose-response curves for Fus3 activation with respect to pheromone. The model displays similar behavior to that observed experimentally, although the experimental curves do not completely agree with one another [37, 38]. Also note that the level of noise observed in our simulations is equivalent to, if not less than, that observed *in vivo*. (B) G-protein activation time-course curves in response to 100nM pheromone. An initial spike in activation, a subsequent decline and a long-term increase are seen upon pheromone stimulation in both wild-type FRET experiments [54] (black circles) and simulation (red solid line). Addition of cycloheximide in the experimental data (black triangles) indicates that the long-term increase in G-protein activation is due to pheromone-induced transcription [54].

where C_i represents the set of unique complexes in simulated cell i, $\triangle$ and $\cup$ are the symmetric difference and union set operators, respectively, and $|X|$ is the cardinality of set X. Given the complexes present in two simulated cells, drift is the number of complexes unique to either one cell or the other, divided by the total number of complexes in the union of the two cells. Drift can thus be interpreted as the probability that a complex found in one cell is not found in the other at a particular point in time. For example, $d = 0$ indicates identical sets of complexes, whereas $d = 1$ means the sets are pairwise disjoint. We performed this comparison between multiple simulation replicates that started from exactly the same steady-state initial condition; thus $d = 0$ at $t = 0$ for all of our simulations (Fig. 2.3A; Sections A.2.3 and A.3.3). Note that this calculation takes into account any difference between complexes, whether the difference is in binding partners, phosphorylation states, or otherwise. Analysis of other potential criteria for differentiating complexes yielded similar results to those discussed below (Fig. A.9 and Section A.3.3).

We observed a marked increase of drift between simulations with pheromone (and thus signaling activity) as opposed to those without pheromone (Fig. 2.3B). At peak Fus3 signaling activity ($t = 360$ seconds), around 80% of all unique complexes were exclusive to one simulation or the other (Fig. 2.3B). Such small overlap indicates that individual cells utilize different sets of signaling complexes, consistent with the ensemble hypothesis [11, 12]. To confirm that this high level of drift is not an artifact of our chosen parameters, we generated over 1000 rule sets with randomized rate parameters (Section A.2.4). In Fig. 2.3C we see the distributions of drift values among scaffold-based signaling species for both the validated model and models with randomized parameters at peak Fus3 signaling. Although the average random parameter set has somewhat lower drift than observed in our original parameter set, approximately 97% of the drift values from the models with randomized parameters were nonetheless greater than 0.8. The high level of drift among signaling species thus likely arises from the rules and interactions themselves rather than specific rate constants.

While the results in Fig. 2.3C indicate relatively high levels of heterogeneity at a particular time point, it could be that two different simulated cells utilize the same set of complexes, just at

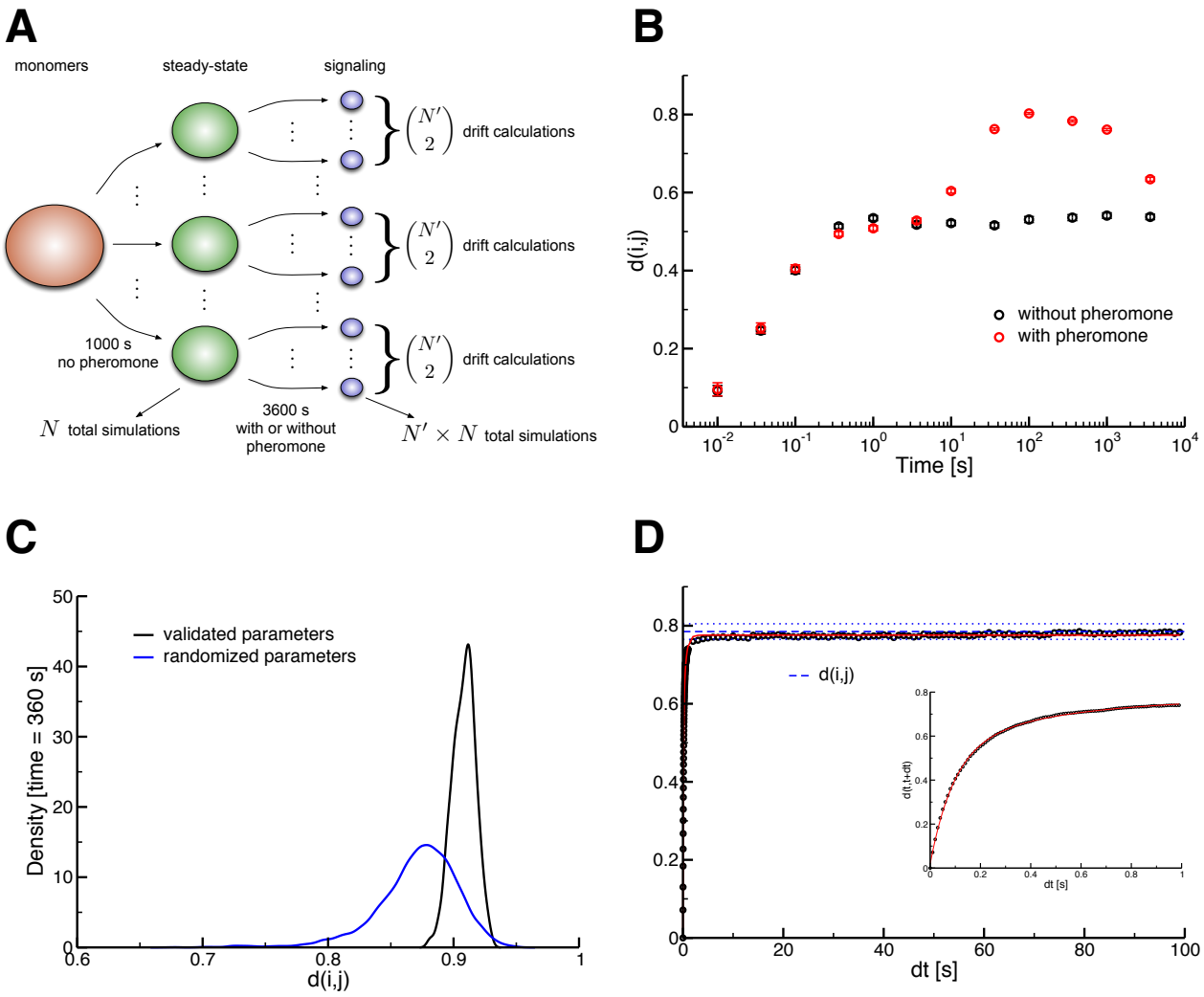


Figure 2.3: Characterization of heterogeneity among signaling species in the ensemble model. All error bars are 95% confidence intervals. (A) Visual depiction of our simulation method. An initial state was defined (red) and a number of trajectories were simulated to represent a set of homeostatic yeast cells (green). Pheromone was added and each steady-state cell was simulated for a number of independent 1-hour trajectories (blue). Drift values were then calculated pairwise for all simulations that were derived from the same original homeostatic cell. The same procedure was also performed without adding pheromone as a control. (B) Average drift on a log time scale ($n = 45$) between simulations starting from the same initial conditions, with peak drift occuring between 100 to 300 seconds. The decline in drift ($t = 3600$ s) is the MAPK network returning to homeostasis. (C) Density of drift (calculated using kernel density estimation; Section A.2.4) among scaffold-based complexes with randomized parameters (black, $n = 9000$) and the parameters that reproduce the experimental data (blue, $n = 450$). (D) Autodrift occurs on two different timescales as indicated by statistical analysis of our double-exponential fit ($n = 10$, Section A.3.3). Signaling events induce rapid divergence within one second (inset) from identical initial states within a particular simulation starting at the time of peak signal output ($t = 300$ s, Section A.3.3) [64]. The blue dashed line is the average drift between simulations at the 360 second time point.

different times during signal transduction. We thus considered the differences between cells based on the union of all the unique complexes they sampled across the time points in our simulations (*i.e.* the points in Fig. 2.3B). We found that using the union of complexes across times only reduced absolute drift levels by about 10%, indicating a high degree of diversity between simulated cells across the entirety of the signaling dynamics (Fig. A.10).

Our analysis of drift across time points raised the question of whether an individual simulation i maintains a specific set of complexes, or if the set changes over time. To answer this question, we used an alternative drift calculation, termed autodrift: $d_i(t, t+\Delta t)$ instead of $d(i, j)$. We found that simulated cells employ rapidly changing sets of complexes during peak signaling times in this model (Fig. 2.3D). Autodrift increased as a double exponential, with a longest time scale of approximately 0.5 seconds (Fig. 2.3D, inset, and Section A.3.3). Indeed, within 5 seconds the difference between a cell and its past self achieves levels of drift similar to that observed between two completely independent cells in the population. This is consistent with observations from both modeling and experimental studies of epidermal growth factor signaling in mammals, where a diverse set of phosphorylated species forms rapidly during signaling [64, 65]. The rapid increase in drift also highlights the transient nature of the ensembles of complexes that are generated.

2.2.4 Detailed analysis of signaling species

It is possible that the putative ensembles in this case merely represent a set of highly similar (though technically distinct) signaling species that form around a large "core" signaling complex. We thus examined in detail the structures of the scaffold-based species at various time points in our simulations. If a core complex were present, we would expect to see substantial conservation of protein binding patterns (ignoring phosphorylation state) in the set of unique complexes. Though Ste5 dimers are present in $\sim 70\%$ of species during peak signal throughput, conservation significantly declines as the binding pattern is expanded to include more proteins (Fig. 2.4A). In fact, not once did we find a Ste5 dimer bound to all its potential interaction partners, indicating that the complex used in the standard graphical depiction of this phosphorylation cascade is one that would very

rarely, if ever, occur in simulations of this model (Fig. 2.1A) [21, 22, 34–38].

It is possible that complexes in the ensemble model still assemble around a consistent core structure, just not the traditional representation of a scaffold-based core signaling complex that we intuitively expect (Fig. 2.1A). Since there are over 3 billion possible scaffold signaling structures in this model, however, we could not search for this core by enumerating all possibilities and looking at conservation patterns as in Fig. 2.4A. We thus used a straightforward clustering analysis to search for an alternative core structure. The signaling species generated in our model were clustered on the basis of the structural similarity between complexes, represented in this case by the *graph edit distance* metric, which is simply the number of changes (or edits) that would be required to form one complex starting from another. This distance accounts for differences in the members of a complex (*i.e.* the removal of a protein from a complex increases the distance) as well as differences in phosphorylation state, etc. (Fig. A.12 and Section A.3.4).

We implemented a hierarchical clustering algorithm based on this distance. Briefly, the algorithm chooses a representative complex from each cluster, called the "clustroid," which is the complex with the lowest average graph edit distance to all other complexes in its cluster (Section A.3.4). At each level of the hierarchy, the algorithm combines the two clusters whose clustroids are most similar, that is those with the minimum graph edit distance (*i.e.* the minimum between-cluster distance, or MBCD). This algorithm is initialized with each complex in its own cluster (meaning the complex is its own clustroid) and continues until the original set of complexes is partitioned into a given number of clusters. This number, which we call the "cutoff," is a free parameter and is relatively arbitrary in our case (Fig. A.15 and Section A.3.4), so we repeated the clustering algorithm with numerous different cutoff values. We calculated the size of the largest conserved structural pattern as a function of the cutoff value for each cluster that contained ten or more complexes. We found that, on average, this conserved pattern contained fewer than 2 proteins (Fig. 2.4B), indicating substantial dissimilarity among clustered proteins; cutoff values producing clusters with 4 or more proteins in the conserved subgraph were very rare (Fig. A.15). These results, combined with the dissimilarity between clusters generated from independent simulations

(Fig. A.13) and the high levels of drift we observe (Fig. 2.3B-D), underscore the strong ensemble character of this model.

2.2.5 Building a machine model based on a multi-subunit kinase

The findings described above indicate that heterogeneous ensembles of complexes can indeed transmit and process extracellular information with levels of noise comparable to those observed experimentally (Figs. 2.2-2.4). To understand if machine-like complexes could also produce reliable signaling behavior, we constructed an alternative model with the goal of assembling signaling machines, which we defined to be stable, multi-subunit kinases based around the scaffold Ste5 [3, 22, 32]. Specifically, the machine we focused on consists of a Ste5 dimer, with each scaffold protein bound to a Ste4-Ste20 dimer and two kinases, Ste11 and Ste7 (Fig. 2.1A). Upon assembly and activation, this decameric structure binds and phosphorylates Fus3 according to standard mass-action kinetics [22]. In contrast to the previous model, we were forced to introduce *a priori* assumptions (neither experimentally supported nor specifically refuted) in order to generate stable signaling machines. The simplest possible approach would be to create rules and rates that render the desired machine complex incredibly stable. The decamer, however, is essentially never generated in our original model's simulations (Fig. 2.4A), so a machine model based purely on increasing the stability of the desired complex is unlikely to actually produce such machines in high quantities reliably. As mentioned above, this fact resembles the Levinthal paradox in protein folding: no matter how stable the native state of a polypeptide chain may be, proteins would essentially never fold if they randomly searched for this state on an otherwise "flat" energy landscape [41, 42]. Alternatively, evidence suggests that molecular machines assemble hierarchically *in vivo* [66], and so we added specific rules that determine the order in which binding and phosphorylation could occur between the scaffold and its associated proteins (Fig. 2.1B, red arrows). This represents a hierarchical energy landscape (extending the analogy to protein folding), where each consecutive step builds toward the formation of a "native" signaling machine [41]. For example, in the machine model, binding of Ste11 to the scaffold can take place only if Ste5 has dimerized

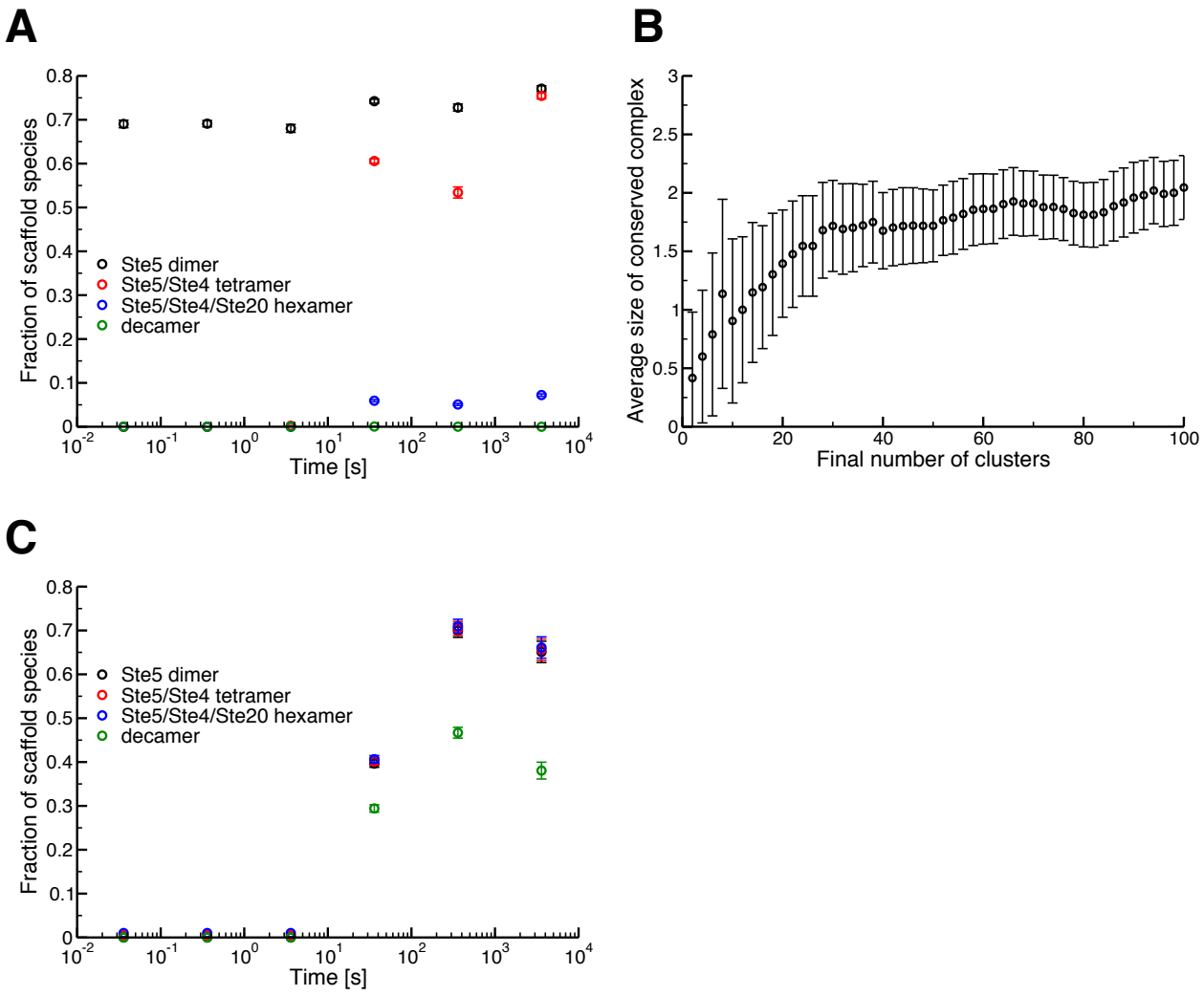


Figure 2.4: Structural analysis of complexes. (A) Structural conservation among scaffold-based signaling species in the ensemble model ($n = 10$). Ste5 dimers are present in 70% of species (black), however as we consider higher-order oligomers formed around this dimer, the fraction of species that contain these patterns drops sharply, with the fully bound Ste5-based decamer not seen at all. The standard depiction of the scaffold-based signaling complex (Fig. 2.1A) is thus unlikely to be observed in the ensemble model, if it occurs at all. Error bars are 95% confidence intervals. (B) The average size of the conserved protein complex as a function of the final number of clusters (*i.e.* the cutoff). At each cutoff, the average considers only those clusters with 10 or more members, to avoid contribution from very small collections of complexes. The average size does not exceed two, even considering up to 100 unique clusters. Error bars represent the 95% confidence interval of the mean. (C) Conservation of structure as in Fig. 2.4A, but in the machine model ($n = 10$). Here the decamer (signaling machine) is present in about 50% of species during peak signaling. The dimer, tetramer and hexamer patterns (black, red and blue, resp.) are present in identical fractions of the unique species, but are separated slightly in the graph for clarity.

and each scaffold is bound to a Ste4-Ste20 dimer. Beyond these scaffold assembly rules, no other alterations were made to the model.

The resulting rule set is sufficiently complex that it is impossible to directly estimate the number of unique species that the machine model could form. We thus translated this model from Kappa into BNGL and used available BioNetGen tools to calculate the total number of species for this rule set [14]. This analysis indicated that the machine model could generate only a total of 1106 possible scaffold-based structures, a decrease of over 6 orders of magnitude compared to the ensemble model (Section A.3.5). The hierarchical assembly rules in this case thus drastically constrain the set of possible species that the model can sample.

2.2.6 Differences between the machine and ensemble models

As with our original model, we subjected this alternative machine model to parameter variation and confirmed that it can reproduce experimental data (Figs. A.6-A.8 and Sections A.1.8 and A.3.2). Although the dose-response and time-course trends of the machine and ensemble models are similar, they exhibit significantly different sets of signaling complexes. As expected, nearly half of all unique scaffold species in the machine model contained the decamer defined above (Fig. 2.4C), indicating wide conservation of the desired core signaling complex, in contrast to the complete lack of conservation observed in the ensemble model (Figs. 2.4A and 2.4B).

The set of species sampled in the machine model also differed dramatically from those produced by the ensemble model. As a gross estimate of this difference, we considered the *cumulative* number of unique scaffold-based species obtained by a set of simulations; that is, the total number of unique complexes that are found in a group of N simulated cells. In the machine model, this number rapidly approaches a maximum value as N increases, saturating at around 800 after considering only 100 simulations (Fig. 2.5A). The machine model thus samples about 70% of the 1106 possible scaffold complexes in a population of ~ 100 cells. The behavior of the ensemble model is strikingly different, sampling a set of unique structures that is nearly two orders of magnitude greater than the machine model (approximately 70,000, Fig. 2.5A), and failing to saturate

even after considering a population of 600 simulated cells. Although the total number of sampled species across these 600 cells is large, it is only 0.0022% of the 3 billion species the ensemble model could theoretically generate.

As one might expect given the results of Fig. 2.5A, we observed large differences in drift during peak signal output between the two models. On average, only 55% of unique scaffold complexes were exclusive to one of two simulations in the machine model, as opposed to 90% in the ensemble model (Fig. 2.5B). As with the ensemble model, we generated 1000 alternative machine models with randomized parameter sets to determine if the level of drift in this case was an artifact of the parameterization of the model. Though the distribution of drift values was fairly wide across these randomized models, in every case we observed considerably less drift than for the validated or randomized ensemble model (Fig. 2.5B). The rules underlying the machine model thus robustly produce dynamics that one might expect for well-established molecular machines like the ribosome or proteasome: a stable, heavily populated core structure with residual diversity arising from assembly intermediates and the association of substrates and/or regulatory factors.

2.2.7 Evaluating experimental evidence for ensembles

Since these two models can both reproduce general pheromone-dependent trends, one might ask if it is possible to differentiate machine- and ensemble-like signaling processes directly using available experimental techniques. The most natural approach would be tandem affinity purification in conjunction with mass spectrometry (TAP/MS), which is widely employed as a high-throughput assay for the discovery and analysis of protein complexes [19]. For example, Gavin *et al*. employed a "socio-affinity" (SA) index designed to extrapolate binary TAP/MS interaction data in order to discover novel "eukaryotic cellular machines" via clustering analysis [19]. To determine whether this technique could discern the nature of *in vivo* signaling complexes, we characterized the signaling species generated in both the ensemble and the machine models using the SA index [19]. There is a high correlation between the SA scores produced from our two models' sets of species (Fig. 2.6A); clustering these scores using the commonly employed MCL algorithm [19, 46, 67, 68]

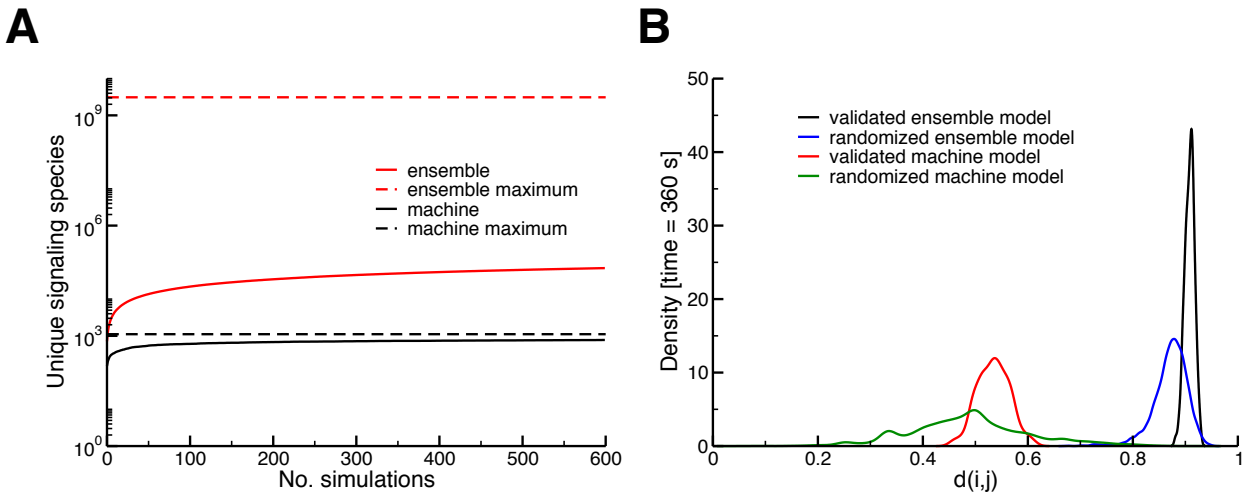


Figure 2.5: Comparison of notable characteristics in the machine and ensemble models. (A) Cumulative number of unique signaling complexes sampled by the machine and ensemble models (black and red, respectively) as a function of the number of independent simulations considered. We see that ensemble model generates a set of complexes approximately two orders of magnitude larger than the machine model over a range of 600 simulations. (B) Drift density among scaffold species in the validated machine model ($n = 450$) and 1000 machine models with randomized parameters ($n = 7789$) as compared to the data in Fig. 2.3C. A large difference between the machine and ensemble models can be seen, with significantly higher mean drift in both the validated ensemble model and the set of randomized ensemble models (both comparisons have a significance of $p < 10^{-5}$, see Section A.2.4). It is also notable that the largest drift value from the machine model is much lower than the smallest from the set of ensemble models with randomized parameters. The remaining heterogeneity observed in the machine model can be attributed to the presence of assembly intermediates and regulatory interactions.

results in essentially the same set of complexes (Fig. 2.6A, inset).

This leads to the question of whether one could ever detect any functional differences between ensembles and machines in a signaling context. Previous work has established the presence of "combinatorial inhibition" [20] (akin to the "prozone" effect [69]) in this particular cascade; increased expression of the Ste5 scaffold leads to a maximal response, past which further overexpression leads to a decline in signal output [21, 57]. We found that the ensemble model reproduces this behavior, while the machine model does not (Fig. 2.6B). In the ensemble model, the eventual decrease in signal response arises because the high quantity of scaffold proteins lowers the probability of cascade components (say, Ste7 and Ste11) binding the same scaffold dimer [20, 69], and so the rate of signal propagation is drastically reduced. The hierarchical assembly rules in the machine model, however, reduce drift by ensuring scaffold dimers can only bind Ste7 after Ste11 is already bound. Beyond a certain minimal point, increasing Ste5 concentration has no effect, since the only potential scaffold binding partners for Ste7 are already bound to Ste11, and thus can propagate signal.

To test if the difference in Fig. 2.6B was robust to variations in the rate parameters, we simulated 100 randomized ensemble models and 100 randomized machine models with three values of Ste5 concentration: Wild Type (WT), 12 times WT (12x) and 60 times WT (60x). We used these simulations to calculate the relative change in peak Fus3 activation (ΔFus3pp) between two pairs of scaffold concentrations: WT to 12x, and 12x to 60x. The validated ensemble and machine models both exhibit a positive ΔFus3pp (12x - WT), corresponding to an increase in Fus3 activation (the peak in Fig. 2.6B); all the randomized ensemble models, and most of the randomized machine models, displayed this same behavior (Figs. A.16, A.17 and Section A.3.7). In the ensemble model, increasing Ste5 to 60x WT concentration decreases response, yielding a negative ΔFus3pp (60x - 12x), while the machine model exhibits an approximately constant response across these concentrations (Figs. 6B and C). The randomized ensemble models also universally showed a decrease in Fus3 activation from 12x to 60x Ste5 concentration, indicating that combinatorial inhibition is a robust feature of the ensemble model. The randomized machine models, however,

had mostly increases in Fus3 activation between these two concentrations, and in no case did we observe a decrease as large as that observed for the ensemble models (Fig. 2.6C). The relative lack of combinatorial inhibition in the machine model is thus likely a feature of the rules themselves, rather than the specific parameters chosen.

It should be noted that the machine considered here is an acyclic complex; that is, there are no ring-like motifs in the protein interaction map for Ste5 (Fig. 1A) [69–73]. Previous modeling studies indicate that ring-like structures can assemble efficiently into well-defined quaternary structures, at least in certain parameter regimes [73]. Nonetheless, overexpression of a single subunit in a heteromeric ring causes a marked decrease in the concentration of the assembled machine, indicating that ring-like structures can simultaneously exhibit a machine-like character and combinatorial inhibition [69, 72, 73]. We leave full consideration of the interplay between robustness and topology in the evolution hierarchical assembly pathways to future work [72, 73].

2.3 Discussion

The nature of the signaling complexes formed during signal transduction is foundational to how we conceptualize and understand information processing in cells. This is particularly true of scaffolds, whose primary function is to serve as a platform for the formation of multicomponent complexes that transmit signals [22]. The question of whether these complexes align more with the machine or ensemble paradigm is thus crucial for developing a broader perspective of the roles scaffolds play. For instance, it has been posited that Ste5 acts to insulate pheromone signals from activating other, related MAP kinase cascades by sequestering active Ste11 in a pheromone-specific complex. This view is inconsistent with the ensembles we observe, however, since those involve appreciable concentrations of free, active Ste11; in contrast, the machine model produces essentially no active Ste11 molecules that are not bound to the scaffold. The capacity of Ste5 to fulfill the role of insulator in this pathway, or the need to posit other mechanisms such as cross-inhibition [22, 34], is thus directly related to the degree of ensemble character the network displays, a fact that high-

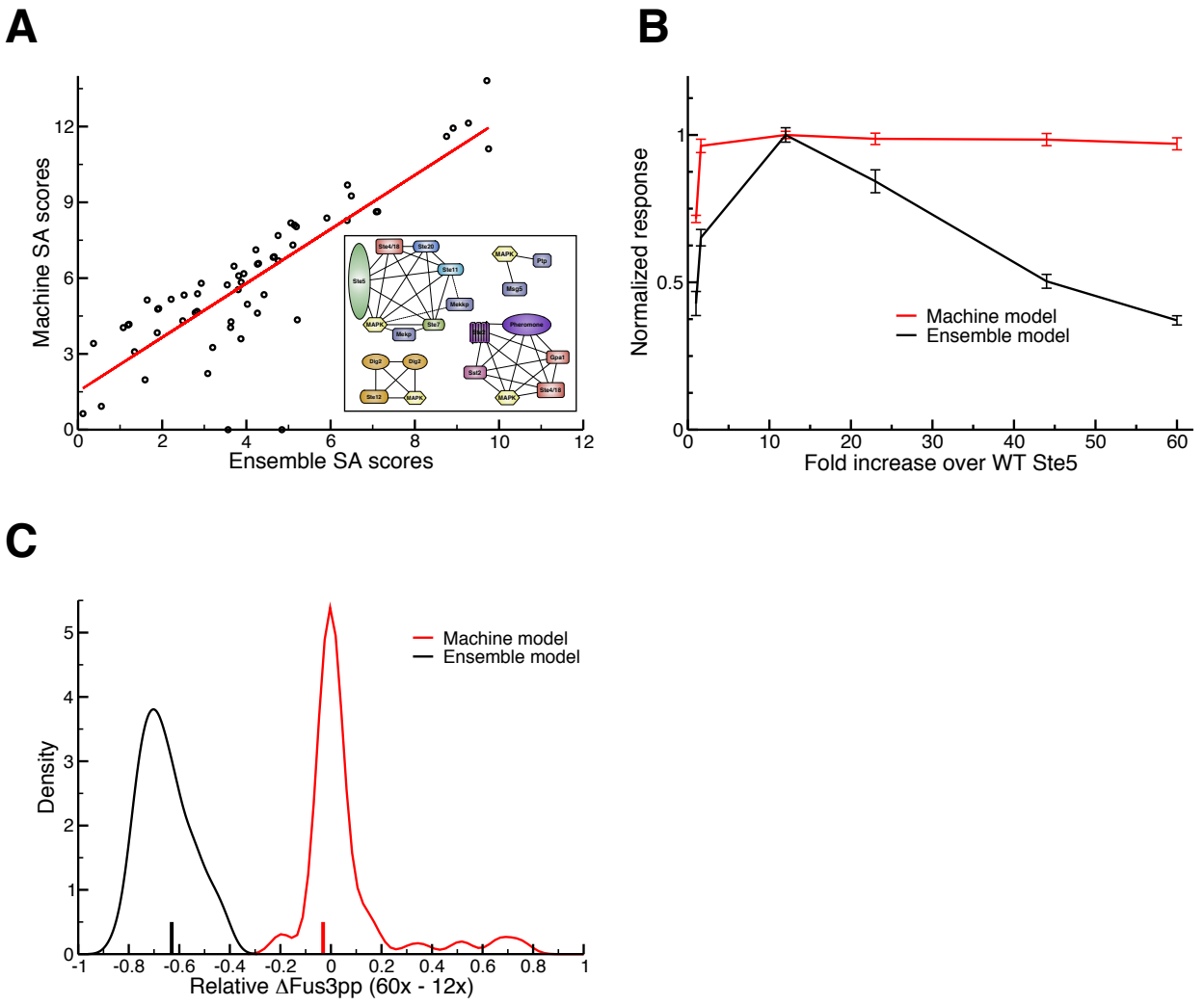


Figure 2.6: Indirect evidence for complex structure. (A) The socio-affinity (SA) scores [19] obtained from computational TAP/MS experiments ($n = 10$) from the ensemble (x-axis) and machine models (y-axis). The machine and ensemble models' scores exhibit a strong correlation and clustering them [46] resulted in highly similar "complexes" (inset, dashed lines exclusive to ensemble model, dotted to machine model), indicating that TAP/MS cannot distinguish between these two modes of signaling. (B) Overexpression of Ste5 in the machine (red) and ensemble (black) model results in different phenotypes ($n = 10$), observing combinatorial inhibition [20] in the ensemble (but not machine) model. (C) We analyzed the robustness of combinatorial inhibition to variations in the parameters by considering 100 randomized ensemble and machine models, each. In each case, we simulated the model with 12x and 60x the WT Ste5 concentration, and calculated the relative change in Fus3 activation, ΔFus3pp (60x - 12x). The negative values for the distribution of ensemble models (black) indicates the robust presence of combinatorial inhibition, whereas the machine models (red) mostly have changes around zero with a few strongly positive outliers (which have been omitted for clarity). The difference in means between the two distributions is statistically significant ($p < 10^{-5}$, permutation test). Vertical bars on the x-axis indicate the relative ΔFus3pp (60x - 12x) for the validated machine and ensemble models in red and black, respectively.

lights the central role that reasoning about quaternary structure plays in developing and evaluating hypothetical signaling mechanisms.

Our findings indicate that certain experimental methods, such as TAP/MS, are ill-equipped to directly resolve the structural details of signaling complexes in living cells. The difficulty in this case lies with the inherently binary nature of co-purification assays: they can tell us that two proteins interact in some way, but they tell us very little about the global structural context of the complexes in which those proteins are found. For example, in our computational TAP/MS experiment, we see that the overall pattern obtained by "tagging" each protein and recording its interaction partners is essentially the same for both the ensemble and machine models (Fig. 2.6A). This is due to the fact that, while the types of quaternary structures formed varies considerably between the two models (Fig. 2.4), the probability of observing any given pairwise association between two proteins is essentially the same. Our results thus indicate that it is problematic to construe clusters obtained from TAP/MS data as representing "cellular machines" in the classic sense [3, 19].

In contrast, experimental methods that can capture ternary or higher interactions (*i.e.* the simultaneous association of three or more distinct proteins) could be used to provide direct evidence for (or against) the hierarchical assembly of a signaling machine. For instance, in the machine model, Ste7 only binds Ste5 after Ste11 is already bound. Observation of Ste7-Ste5 association in the absence of Ste11 binding to Ste5 would thus provide evidence against the type of signaling machine considered here (Fig. 2.1B). Methods such as fragment complementation assays and fluorescence triple correlation spectroscopy could likely be used to probe these types of ternary association dynamics [46–48]. Alternatively, recent advances in single-molecule (super-resolution) microscopy (*e.g.* methods like PALM and STORM) could potentially track the assembly of machine- or ensemble-like signaling complexes [50–53].

While direct experimental tests of the ensemble hypothesis are currently lacking, inherent functional differences between machine and ensemble models can be used to provide indirect evidence for or against a particular paradigm. For instance, the hierarchical assembly rules that are required

to reliably construct a functional scaffold-based signaling machine prevent our machine model from replicating the experimental observation of combinatorial inhibition (Fig. 2.6B) [20–22]. Our analysis of machine models with randomized parameters indicate that this is likely a general observation: in order to exhibit combinatorial inhibition, signaling networks must have the capacity to sample large sets of complexes, ultimately leading to ensemble behavior (Fig. 2.6C). Although more work is clearly needed to unambiguously resolve the question of machines vs. ensembles, our findings on combinatorial inhibition indicate that at least some degree of ensemble character is likely present in yeast pheromone signaling. It is also clear that the assembly pathways employed to form machines can have measurable, phenotypic consequences. As a result, even if one could determine experimentally the small set of machine-like complexes employed by some network, making a model that employs these machines, but ignores the mechanisms necessary to generate them [36, 43], may not accurately capture the response of the system to perturbations.

The presence of ensemble character in signaling also highlights a potential evolutionary trade-off between machines and ensembles in terms of their phenotypic plasticity. Considering again the analogy to protein folding, adopting a well defined, thermodynamically stable tertiary structure clearly enables the function of a vast array of protein domains (*i.e.* the general protein structure-function paradigm) [74]. In some cases, however, it has been posited that "intrinsically unstructured" (or unfolded) protein domains may have a distinct functional or evolutionary advantage: for instance, they may display greater interaction plasticity, binding specifically yet transiently with a large number of protein targets [74, 75]. Similarly, a protein with a robust, stable quaternary structure (*i.e.* a machine) [3, 11, 12] may be beneficial for the conservation of universal cellular tasks, like protein synthesis and degradation. In the case of signal transduction, however, ensembles may offer greater functional and evolutionary plasticity. For example, modifying Ste5 expression levels produces altered, but nonetheless functional, responses without the need to introduce complex, coordinated mutations to the reaction network's rule set (Fig. 2.6B) [25]. In this sense, both intrinsically disordered proteins and pleiomorphic ensembles may perform unique intracellular tasks precisely because they involve less well-ordered (tertiary or quaternary) structures. The ensemble

character we observe could thus represent a form of weak regulatory linkage among genes, ultimately being responsible for the remarkable capacity of MAPK networks to exhibit different but meaningful phenotypes when they are re-wired, either through synthetic modifications or naturally over the course of evolution [22, 25, 45, 76].

Since machines do indeed form in some signaling networks (*e.g.* the apoptosome), there is likely a spectrum of structural specificity in the formation of complexes during signal transduction [3, 11, 33]. Indeed, one could modify the machine model presented here to include a finite probability of "off-pathway" binding events (*e.g.* some chance that Ste7 will bind Ste5 even if Ste11 is not already bound). Such models could exhibit intermediate levels of both drift and combinatorial inhibition (Figs. 2.5B and 2.6B); future work on this and related systems will be necessary to understand the particular functional and evolutionary consequences of a particular degree of ensemble-like character in any given system. Nonetheless, our work clearly demonstrates that large, heterogeneous ensembles can indeed reliably transmit and interpret extracellular information [11, 42, 43]. This hints at the existence of a new paradigm for molecular computation, one in which the evolution or engineering of "local" interaction rules allows for robust information processing in the absence of "global" order (*i.e.* a stable, multi-subunit signaling machine) [3, 32]. Understanding the consequences of this paradigm for robustness [77], plasticity [22, 25] and crosstalk [34] in signaling networks represents a crucial task for the emerging field of systems biology.

2.4 Methods

2.4.1 Simulation

The models in this work were simulated using KaSim, a stochastic simulator for rule-based models based on the Kappa language that is capable of stochastically sampling all possible species a given model can generate (Fig. 2.5B; Section A.2.1) [23,24]. The model is initialized with a set of (mostly) monomeric protein agents and simulated for 1000 seconds without pheromone to generate a steady-state population of N untreated "cells." We treated the cells with pheromone, and

generated a set of N' independent hour-long simulations from each steady-state starting cell. All of the complexes present in the simulation were recorded at logarithmically spaced time intervals. Compositional drift calculations were performed using these "snapshots;" we only performed this calculation between simulations that started from exactly the same initial conditions (Fig. 2.3A). We performed similar simulations to determine both dose-response and the time course trends. Further simulation details may be found in Section A.2.3.

2.4.2 Autodrift statistical fitting

Time-dependent drift trends were fit to a set of exponential models using nonlinear least-squares regression. We found that a double exponential function was the best fit for the data upon analysis of the residuals and the statistical significance of the estimated model coefficients. The functional form of the model and the full statistical analysis can be found in Section A.3.3.

2.4.3 Complex classification and clustering

We focused primarily on the scaffold-based species for the analysis of structural conservation and subsequent clustering. These were defined as any complex that included a Ste5 agent or that could bind a free Ste5 agent. We created a vector notation to uniquely identify any scaffold-based complex to simplify the calculation of the graph edit distance between any two complexes (Figs. A.11, A.12, and Section A.3.4). We then implemented the clustroid-based hierarchical clustering approach. Other clustering criteria, such as standard single- and complete-linkage, gave similar results (Section A.3.4).

2.4.4 Socio-affinity scores and complex determination

We extracted all the binary interactions from the set of complexes generated by our simulations, artificially creating "bait" and "prey" association data. This computational version of the TAP/MS experimental procedure was used to generate the SA scores [19]. The MCL clustering algorithm

[68] was then employed to generate the “functional modules” generally associated with such data sets [67]. More information on the SA score calculation and clustering algorithm can be found in Section A.3.6.

Chapter 3

Understanding the Dynamics of Scaffold-Mediated Signaling

3.1 Introduction

Intracellular signaling networks form the basis for cellular adaptation to the environment, and scaffold proteins (which serve as nucleation points for the assembly of signaling complexes) are a common component of these networks in eukaryotes. Despite the fact that scaffold proteins have been the targets of numerous studies [20, 55, 57, 78–84], an understanding of the general dynamical behaviors that these multivalent adaptor proteins impart to signal transduction has yet to emerge. Prior work has shown that some specific dynamical features rely on the binding context of effector proteins to scaffolds [1]. However, a comprehensive characterization of the relevant mechanistic details still does not exist even for well-studied networks such as the scaffold-dependent yeast pheromone signaling network [35]. Regardless, a wide range of posited functions for, and consequences of, scaffold proteins in signaling networks have been put forth [1, 11, 20, 22, 23, 80].

In a recent review, Lim and coworkers outlined a number of these hypotheses and the associated intuition by which scaffold proteins might impact signal transduction, some of which are based on their studies of the Ste5 scaffold and most reference at least one specific scaffold [22]. Perhaps the

most ubiquitous is the supposition that scaffolds in kinase cascades prevent signal amplification [22, 23]. The argument for this hypothesis is based on the intuition that stoichiometric limitations imposed by the scaffold somehow limit activation of downstream species [22, 23]. Another such hypothesis states that scaffold proteins should prevent inappropriate activation of pathways (a form of *crosstalk*) via sequestration of shared signaling components [24]. These and other conjectures continue to be reiterated in literature without the appropriate theoretical work to address their feasibility.

In contrast, there have been a select few investigations that have rigorously examined the general influence of multivalent scaffold proteins on signal transduction. Perhaps the most prominent dynamical feature of scaffolding is *combinatorial inhibition*, a phenomenon similar to the *prozone effect* observed in immune response in which excess scaffold concentration inhibits signal throughput [20]. In fact, combinatorial inhibition has been experimentally confirmed in the yeast pheromone signaling network [21], which is composed of a MAPK cascade that is dependent on the presence of a scaffold [21, 85]. Another prominent model-driven investigation explored how scaffold proteins influence signaling dynamics using spatial stochastic simulation [80]. Locasale, *et al*. showed that scaffolds are capable of preventing signal attenuation in kinase cascades that exhibit strong phosphatase activity. In their model, increasing affinity between kinases and the scaffold correspondingly increases the probability of activation events, since kinases are more likely to be located on a scaffold. However, this study has a few key limitations, including the fact that enzymes are incapable of saturation since substrate activation is modeled using instantaneous collision events. Finally, a number of studies have examined scaffold-dependent nucleation of signaling species in various contexts. One focuses on optimal assembly of fully-bound scaffold molecules [84] and another on the effects of combinatorial complexity due to independently operating binding sites on scaffold proteins [1]. In the latter study, distinct mechanisms of scaffold-based signaling complex assembly were found to generate unique scaffold-dependent phenotypes.

In this work, we expand and improve upon these prior studies to construct a framework with which to systematically investigate a number of these posited functions of scaffold proteins purely

as a result of their multivalency. Unfortunately, as previously mentioned, the mechanistic details with which signaling proteins assemble on the scaffold remain poorly characterized [12], yet constructing models of signal transduction requires this knowledge. With a model of the yeast pheromone signaling system, we recently demonstrated that signaling networks can function using strongly contrasting assembly strategies [1]. Both heterogeneous *ensembles* of protein complexes that were posited by Mayer *et al*. to be capable of reliable signal transduction [11], and discrete, well-organized signaling machines [1] can replicate most experimentally observed signaling dynamics, however only ensemble signaling could recapitulate the experimentally observed combinatorial inhibition [1, 21]. Beyond this, however, there is currently no general understanding of how ensemble-like scaffold dynamics might differ from the machine-like case, or how both of those compare to signaling cascades that are not based on a scaffold. Therefore, we created models of all three scenarios (ensemble, machine and a "solution" case, with no scaffold at all), and characterized the relevant phenotypic properties (*i.e*. dose-response trends, time to steady-state, noise suppression, *etc*.) while varying key aspects of the system (*e.g.* scaffold valency and concentration). We found that, in some cases, the mere presence of a scaffold protein, independent of the assembly paradigm, influences the dynamical properties of the network (*e.g*. the variability in response), whereas in other cases, behavior is paradigm-dependent (*e.g*. robustness to crosstalk). However, the most striking result is the existence of strong signal amplification in all three models, with the machine model exhibiting the highest amplification. This stands in contrast to the long-standing hypothesis that scaffolds generally prevent signal amplification [22, 23].

This research establishes a basic understanding of how various scaffold-based signaling paradigms differ from one another, and how various evolutionary pressures (*e.g*. a pressure for fast responses, or a pressure for robustness to fluctuations in scaffold protein concentrations) might influence how a multivalent scaffold protein is employed in any particular system. Our findings also provide a basis for designing experiments aimed at understanding which particular scaffold-based assembly paradigm is employed in any specific signaling cascade. Furthermore, these findings have important implications for synthetic biology in particular, where the goal is to manipulate the intracellular

dynamics of response to signal in order to produce highly specific responses [22, 78]. Most importantly, however, this work highlights the necessity of theoretical, model-driven research as a means of confirming or disproving the viability of seemingly intuitive hypotheses in dynamical systems biology.

3.2 Results

3.2.1 Model construction

As a brief recap of the ensemble and machine signaling paradigms (described in Suderman & Deeds [1]), our model of ensemble-based signaling, signal transduction events depend purely on local interactions; kinases bind the scaffold independently of one another, and a fully assembled scaffold complex need not be formed for the signal to propagate (Fig. 3.1A, black and red lines) [1, 12]. In contrast to ensemble-like signaling, proteins associate with the scaffold in a particular order within our machine-like signaling models, so that the possible binding reactions are driven by the global state of the complex. In this paradigm, signaling machines are constructed in a hierarchical manner (Fig. 3.1A, only red lines), ultimately forming a multi-subunit enzyme that activates downstream components (*e.g.* transcription factors) only when fully formed [1]. Finally, the solution model operates as a set of independent kinases that directly bind and phosphorylate the next protein in the cascade [86].

The most fundamental aspect of these models' construction is the implementation of the scaffold-kinase binding rules. For all scaffold-based models, we required that signal transduction occurs via scaffold-bound signaling species, compared to prior theoretical investigations in which the signal could propagate regardless of whether the kinases were bound to the scaffold [80]. Our models' scaffold proteins are based on those found in the yeast pheromone MAPK network [88], and thus activation of any kinase in the MAPK cascade cannot occur in the absence of scaffold proteins. Furthermore, since the simulations take place in a well-mixed environment [89], our analyses are solely concerned with how the multivalent nature of scaffolds as adaptor proteins influence the

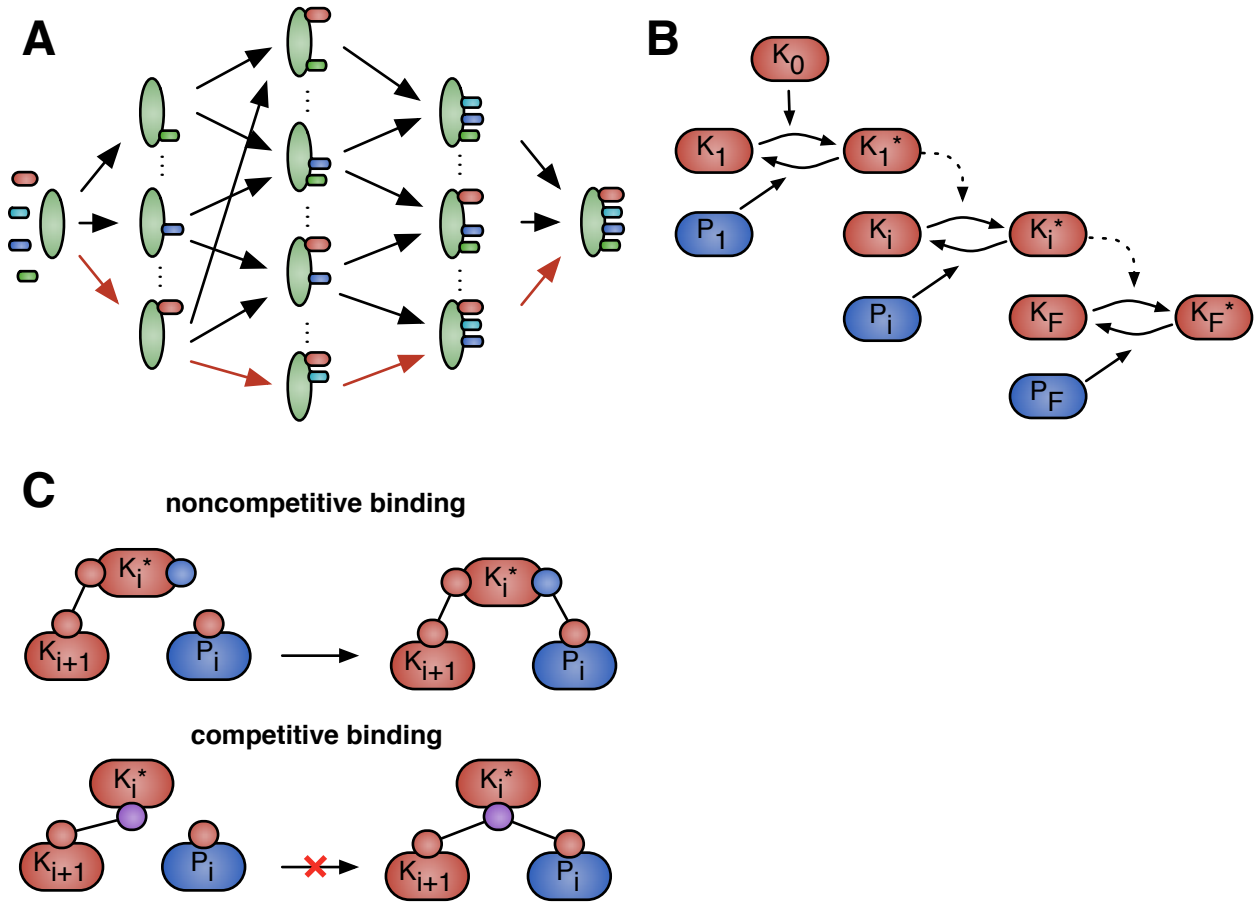


Figure 3.1: Schematics of key interaction types in scaffold-dependent signaling paradigms. Signaling components, *e.g.* kinases (small, variously colored), bind to a scaffold (large, green) in order to propagate signal. Components may either bind independently of the scaffold's binding state (black and red lines) or hierarchically (red lines), representing the ensemble and machine signaling paradigms, respectively [1, 12]. Note that only in the machine signaling paradigm is the right-most complex required; ensemble signaling requires only neighboring components to be simultaneously scaffold-bound for signal propagation. (B) A multi-step kinase cascade based on Goldbeter and Koshland's covalent modification cycle [87]. Here, some kinase (K_i) activates the next kinase in the cascade $(K_{i+1} \rightarrow K*_{i+1})$ and its associated phosphatase (P_i) similarly deactivates it $(K*_i \rightarrow K_i)$. (C) Traditional enzyme kinetics involves competition in binding between a kinase's phosphatase and substrate (*competitive* binding, bottom). However, since the machine and ensemble signaling paradigms allow phosphatases to bind and dephosphorylate kinases both on and off the scaffold, we implemented *noncompetitive* binding behavior (top) in the solution model as a more relevant control. Kinases in the solution models may therefore bind a substrate and phosphatase simultaneously.

dynamics of signaling and not with the spatial effects of scaffolds.

Stimulation of both ensemble- and machine-like cascades takes place via a signaling agent that enzymatically activates the first of a series of N kinases. The strength of activation is determined by modifying the catalytic rate of the first kinase's phosphatase; signal strength is measured as the ratio of the maximum velocities of this first "stimulation" agent and the phosphatase that acts on the first kinase in the cascade (see the Section B.1 for further details) [87]. Each subsequent kinase, which also has a corresponding phosphatase to prevent undue cascade saturation [86], binds the scaffold and propagates signal according to paradigm-specific rules. Our ensemble-like signaling models require only that an active kinase and its substrate are simultaneously bound to the scaffold for phosphorylation to occur; machine-like signaling requires that all upstream association and phosphorylation events have also occurred (Fig. 3.1A). The quantity of activated final kinase (K_F^*) is considered the output of the cascade; our output is thus distinct from previous theoretical studies, where the number of fully-assembled scaffolds in the system was considered to be the output [84]. Previous models of scaffold-based signaling have considered how phosphatase-based dephosphorylation of scaffold-bound kinases is implemented, and found that, in many cases, changes in qualitative signaling behavior are minimal [20, 80]. Due to a lack of evidence to the contrary, we therefore assume that phosphatases may operate on scaffold-bound kinases with the same activity and parameters as freely diffusing (*i.e.* unbound) kinases.

In addition to our two scaffold-based signaling paradigms, we implemented a scaffoldless "solution" model to serve as a control. This multi-stage cascade is based on the covalent modification cycle outlined by Goldbeter & Koshland and extended by Rowland *et al.* (Fig. 3.1B) [86, 87]. Importantly, we modified the typical representation of this process to allow phosphatase-mediated deactivation of substrate-bound kinases. This change reflects the ability of phosphatases in the machine and ensemble paradigms to dephosphorylate scaffold-bound kinases, and thus serves as an additional measure of control for the two scaffold-based models. We label this type of model *noncompetitive* since substrate and phosphatase can simultaneously bind an active kinase (Fig. 3.1C, top). This has the interesting impact of causing the first phosphorylation cycle, or Goldbeter-

Koshland (GK) loop, in the cascade to have the properties of an isolated GK loop because there is no sequestration of the modified substrate (*i.e.* the second kinase) in the subsequent GK loop's kinase-substrate complex (the *competitive* model; Fig. 3.1C, bottom). Said another way, the phosphatase of the initial loop may bind the kinase-substrate complex of the second loop and thus has access to the entire pool of active second kinase.

The kinetic parameters for these models were chosen based on the parameters from a previously outlined model of the yeast pheromone signaling pathway [1]. For simplicity's sake, kinase copy numbers are identical to one another except for the final kinase, which is at a copy number that is 10-fold larger than all other kinases, and interactions between specific protein types (*e.g.* kinase-scaffold or phosphatase-kinase) have identical kinetics. Since the initial rate parameters we chose resulted in enzymes that were universally unsaturated (*i.e.* substrates were always at low concentrations compared to the K_M values of their kinases and phosphatases), we constructed a second set of parameters to consider the influence of enzyme saturation. In these saturated models, the K_M of any arbitrary kinase-substrate pair was at least 2 orders of magnitude smaller than the substrate concentration (Table 3.1). Our results focus mainly on the models acting in the unsaturated parameter regime, since the noncompetitive nature of the phosphatases induces a strong switch-like behavior in the saturated cascades across all three signaling paradigms (see Section B.3).

3.2.2 Steady state dose-response trends

The first step in characterizing these signaling paradigms was to generate sets of dose-response data while varying key aspects of the cascade, namely the phosphatase copy number and the number of distinct kinase types in the cascade (equivalent to the number of kinase binding sites on the scaffold, which we refer to as the cascade's *depth*; Fig. 3.2A). The resultant dose-response trends were universally sigmoidal in shape, and so we characterized the behavior of each model by fitting

the response data to a Hill function:

$$R = R_{\max} \cdot \frac{S^n}{S_{50}^n + S^n}. \tag{3.1}$$

We can thus describe the steady state response properties of each model in terms of the Hill function's parameters: maximum response $(R_{\max})$, sensitivity to signal (S_{50}; the signal producing a half-maximal response), and response ultrasensitivity (n; the sharpness of the switch from minimum to maximum response). A representative data set is shown in Fig. 3.2B with the $R_{\max}$ and S_{50} parameters obtained from the fit indicated. In general, our analyses refer mainly to models with 100 phosphatases for each kinase, so that there is a 1:10 ratio of phosphatases to kinases (except with the final kinase in the cascade where there is a 1:100 ratio) unless otherwise noted. This allows for stronger signal throughput as compared to models with higher phosphatase to kinase ratios.

The Hill parameter governing maximal response, $R_{\max}$, is nearly identical between machine and solution models when both are in the same parameter regime (Fig. 3.2A). For these two paradigms, over 90% of final kinase pool is active at steady state when stimulated with a strong activating signal, regardless of whether the kinases in the cascade are unsaturated or saturated. On the other hand, the unsaturated ensemble models exhibit a much lower maximum response, with about 40% activation of the final kinase concentration even at very high levels of cascade stimulation. This indicates that the maximum response of a network is much more dependent on the rules governing the protein interactions than the presence of a scaffold protein, a trend that is consistent throughout this work. In other words, it is not the mere presence of the scaffold itself, but rather how the scaffold-based signaling complex assembles that ultimately determines $R_{\max}$.

Prior theoretical studies have shown that increasing the depth of a scaffoldless cascade increases the sensitivity to signal (*i.e.* decreases S_{50}) [86, 87]. Our results support this claim (Fig. 3.3A), despite operating in a different parameter regime. Similar to the maximal response trends described above, addition of a scaffold protein alters the quantitative response in a paradigm-

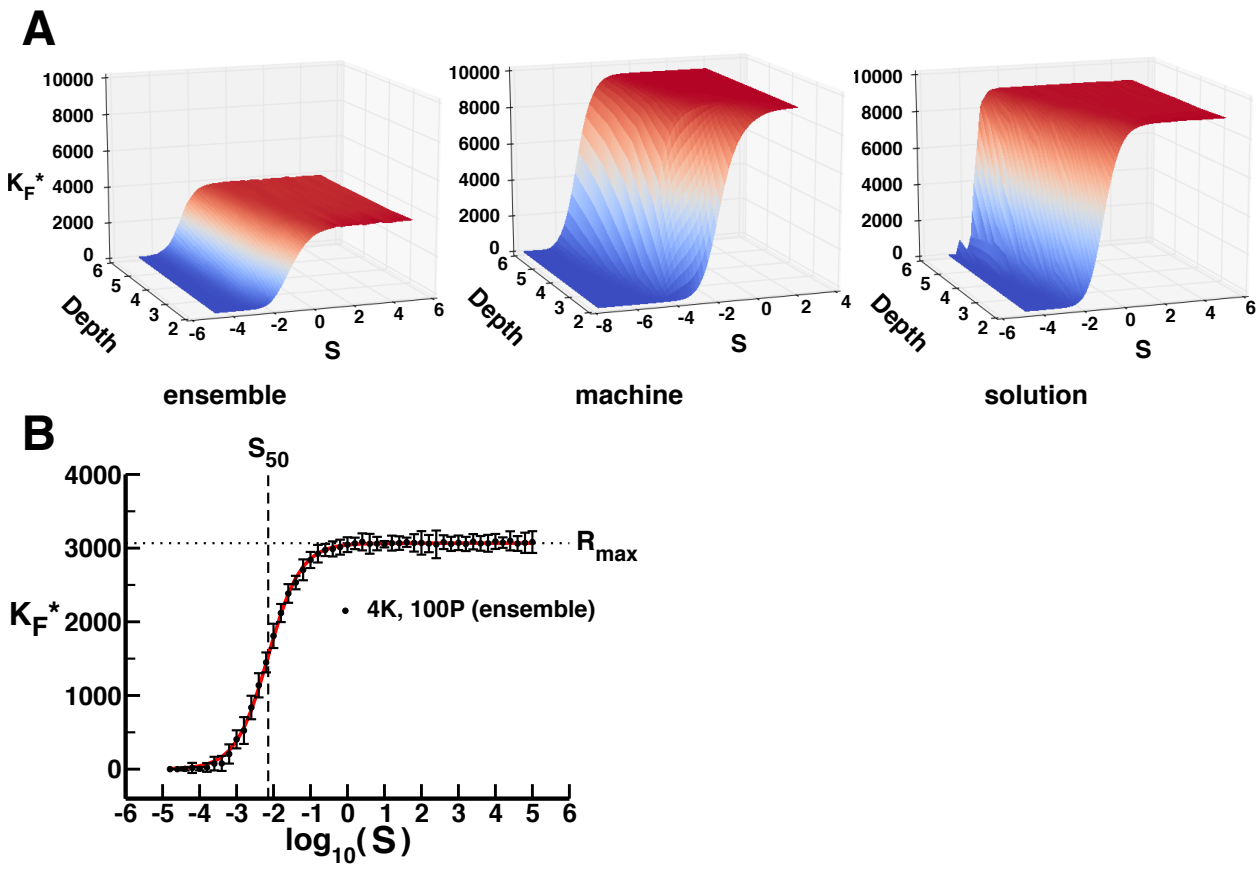


Figure 3.2: Dose-response dynamics for the different signaling paradigms. (A) Dose-response surfaces for unsaturated, low phosphatase simulations of the three signaling paradigms. Depth describes the number of stages in the multi-kinase cascade, S is signal strength, and $K*_F$ is the number of active final kinases, which we consider to be the output. We used simple linear interpolation to smooth these surfaces. (B) A representative data set from ensemble model simulations (10 replications, with 95% confidence intervals about the mean) for a depth of 4 kinases (4K) and a 1:10 phosphatase to kinase ratio (100P). The x- and y-axes (S and K_F^*, respectively) are as in (A). The solid line is the 3-parameter Hill function fit, where $R_{\max} = 3067$ (dotted line), $S_{50} = 0.00717$ (dashed line), and $n = 0.991$; all parameters are statistically significant ($p < 10^{-16}$).

dependent manner. The ensemble models exhibit increased sensitivity to signal as a function of cascade depth, but the increase is shallower than the increase observed for solution models. The increase in sensitivity with cascade depth for the machine models, on the other hand, is much sharper, further highlighting the fact that the knowledge of the binding mechanisms between kinases and scaffold proteins is central to understanding how scaffolds perturb the response to incoming signals. As a side note, the sensitivity of saturated scaffold-based simulations is essentially invariant with respect to cascade depth when the phosphatase-to-kinase ratio is 1:10 (see Section B.3).

This increase in sensitivity to signal with cascade depth directly impacts another posited role in signaling dynamics for scaffolds, which is a mechanism for prevention of signal amplification [22, 23, 80]. Specifically, the hypothesis is that scaffold proteins might limit signal amplification due to stoichiometric constraints on the assembly of relevant signaling species. In order to examine this systematically in our three signaling paradigms, we defined signal amplification similarly to Locasale, *et al*. as the ratio of the final kinase's activity to the first kinase's activity: $\frac{K_F^*}{K_1^*}$. Our results reveal that all signaling paradigms exhibit some degree of signal amplification at moderately low levels of signal (Fig. 3.3B, Section B.2.1). The reason for this, as alluded to above, results from the increased sensitivity corresponding to increased cascade depth (Fig. 3.3A). As the depth of a cascade increases the relatively low signal levels that activate only a small portion of the K1 pool (which behaves as a substrate within an isolated GK loop in all three signaling paradigms) may subsequently activate all final kinase molecules.

Additionally, the presence of modified scaffold proteins in a signaling network has been shown to modify the steepness of the dose-response curve [78]. As a result, we expect that differences in scaffold implementation could impact the steepness or ultrasensitivity of the dose-response curve as characterized by the Hill coefficient, n. Our models show reduced ultrasensitivity for the ensemble and machine paradigms as compared to the solution paradigm in the unsaturated, low phosphatase parameter regime (Fig. 3.3C). Quantitatively, the saturated cascades have a much larger n relative to the unsaturated cascades (as might be expected from prior analyses of GK loops [86, 87]). It is important to note that our simulated data sets lack the signal-space resolution for

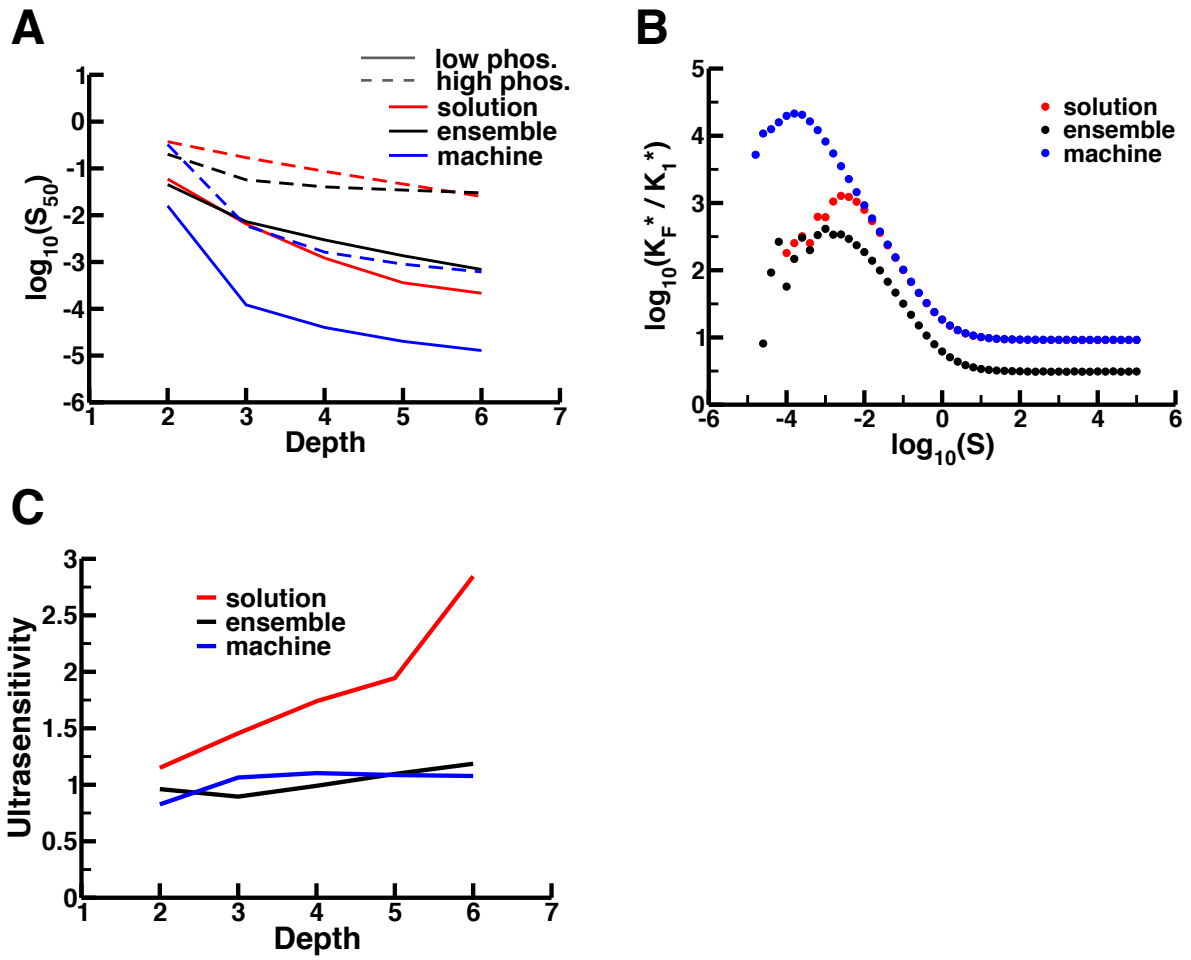


Figure 3.3: Scaffolding generates both general and paradigm-specific behaviors. (A) Examining sensitivity to signal (y-axis) given the depth of the cascade (x-axis) for both high (dashed lines, $P_{i>1} = 1000$) and low (solid lines, $P_{i>1} = 100$) phosphatase activity reveals that lower phosphatase activity, as well as increased cascade depth, leads to increased sensitivity. Notably, for cascades with depth ≤ 3, machine-based signaling generally exhibits increased sensitivity to signal compared to ensemble and solution based signaling regardless of the level of phosphatase activity. (B) Signal amplification, defined as the ratio of first to last kinase activity in a cascade $\left(\frac{K_F^*}{K_1^*}\right)$, occurs in all signaling paradigms. The data shown here are taken from models with depth $= 4$. The underlying cause is a shift in signal sensitivity with cascade depth (A), which induces this amplification at moderately low signal levels. (C) Scaffolding decreases the ultrasensitivity of the response in unsaturated models with low phosphatase activity. Despite the stark difference in scaffold protein assembly in the ensemble and machine paradigms, the scaffold has a similar "linearizing" effect relative to the solution paradigm.

accurate characterization of n for saturated models, since they are all extremely ultrasensitive compared to the unsaturated models. Further simulations would need to be performed to thoroughly characterize the extreme ultrasensitivity of the saturated models, and this computationally expensive task is outside the scope of this study. Nonetheless, previous hypotheses regarding scaffold-induced dose-response linearization are supported by our findings both the machine and ensemble models in the unsaturated regime.

3.2.3 Speed and reliability of response

In addition to steady state dose-response behavior, other properties of signaling networks could easily contribute to their function and evolution. One such property is the speed at which cells are able to respond to some environmental stimulus. We explored the influence that scaffold proteins have on the speed of response by calculating the time it takes for a simulation to reach a response greater than half of that observed at steady state (T_{50}). We calculated this value at two signaling strengths: the signal nearest that required to reach half-maximal response (S^{50}) and the signal resulting in maximal response (S_{max}). In the unsaturated models, T_{50} increases monotonically with cascade depth for all three signaling paradigms at both S_{max} and S_{50} (Figs. 3.4A and 3.4B). However, the machine model consistently takes longer to respond, likely due to the time required to successfully assemble discrete signaling machines on the scaffold. In fact, the machine model does not reach T_{50} for nearly one day of simulated time for signal values nearest S_{50} (Fig. 3.4A). On the other hand, there is negligible difference between the ensemble and solution model response times at both signal values (with the exception of the two-kinase cascades). As observed above, the influence of a scaffold on signaling dynamics in this case is highly dependent on the nature of the binding rules themselves.

In addition to providing a timely response, most signaling networks must reliably respond to signals on the single-cell level (*e.g.* gradient tracking for chemotaxis or shmoo formation in yeast). Reduction of biochemical noise may thus be a key property of signaling cascades, and we posited that scaffold proteins could play a role in controlling fluctuations. To test this possibility, we

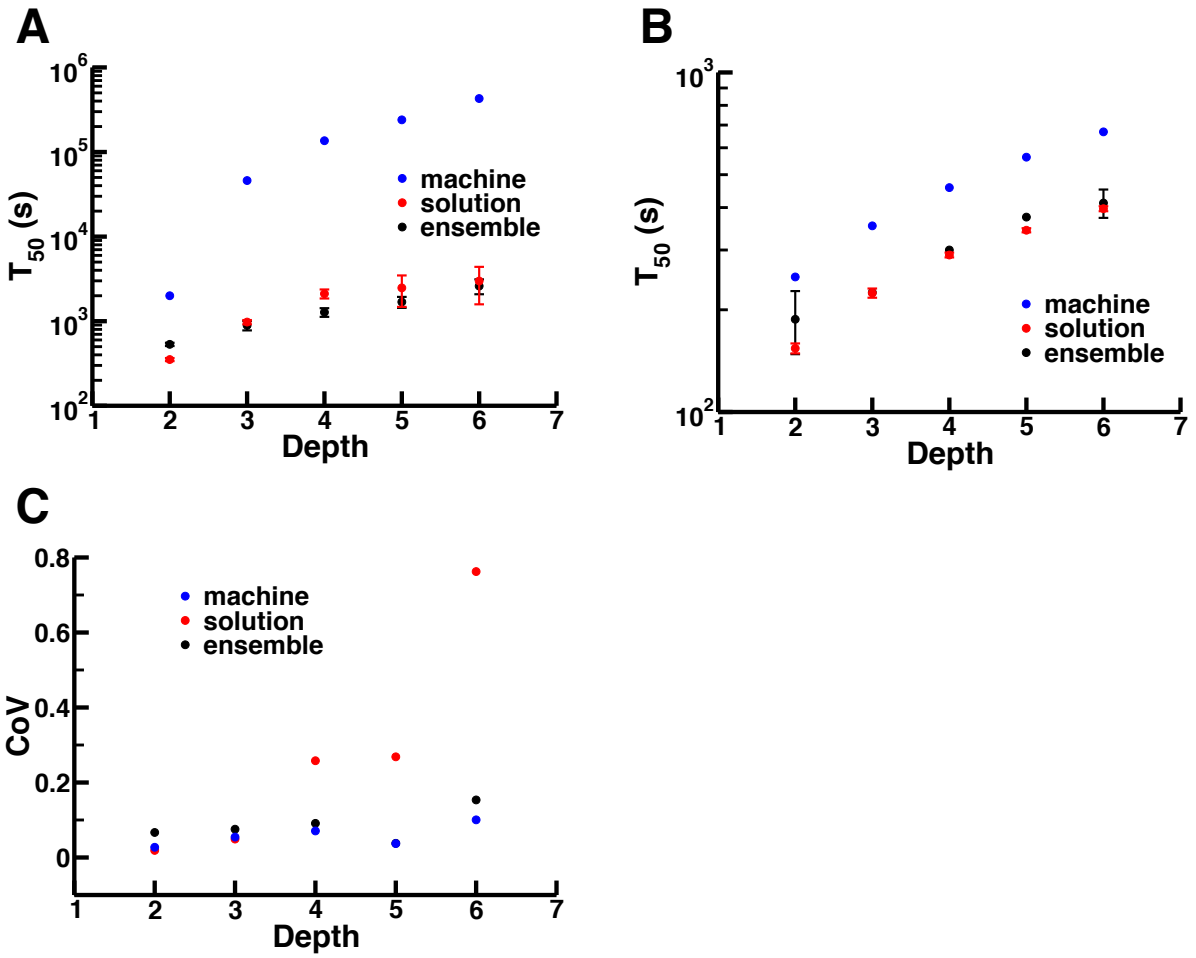


Figure 3.4: Scaffolding alters variability and speed of response. (A & B) Response time at S_{50} (A) and S_{max} (B) increases with cascade depth. These plots indicate the length of time (in simulated seconds) taken to exceed half the observed response at steady state (T_{50}; y-axis) as a function of cascade depth (x-axis). Similar to R_{max} and S_{50} (Figs. 3.2A and 3.3A, respectively), the response time does not specifically depend on the presence of the scaffold, but rather on the assembly paradigm. Strikingly, the machine model exhibits response times nearly two orders of magnitude greater than those observed in ensemble and solution models for deeper cascades with intermediate signal strength (A). (C) Scaffold proteins suppress the noise present in deep solution cascades independent of the assembly paradigm. The coefficients of variation (CoV, y-axis) are taken from the simulation whose signal is nearest the fitted S_{50} value.

examined the variability in response (as measured by the coefficient of variation) for simulations with signal values nearest to their respective S_{50} values and found that scaffolds strongly reduce intrinsic noise for intermediate response values (*i.e.* the steepest region of the dose-response curve), especially for relatively deep cascades (Fig. 3.4C). This makes intuitive sense in the case of the machine model: by constructing a discrete multimeric enzyme (signaling machine) instead of relying on a series of GK loops to activate the final kinase in a cascade, the machine model exhibits less variability in active kinase numbers during signal transduction, thus limiting noise. Perhaps more interesting, however, is the fact that the ensemble models also significantly reduce noise levels, which is particularly striking when considering that the ensemble models are sufficiently combinatorially complex to generate nearly an order of magnitude more signaling species than the machine and solution models in deeper cascades (Section B.4). These results indicate that scaffold proteins provide a mechanism for reducing fluctuations regardless of how they assemble, damping the noise that can arise from the strong response amplification present in cascades that do not utilize a scaffold.

3.2.4 Effects of scaffold number variation

The results described above were all produced with a stoichiometric ratio of scaffold proteins to kinases for kinases 1 through $N-1$ (where N is the cascade depth). It has been shown repeatedly that variations in scaffold concentration can have strong and sometimes nonintuitive effects on network response. The best characterized of these is the *prozone* effect, or *combinatorial inhibition* [20, 21]. We therefore examined our models in the context of scaffold copy number, varying this quantity by over two orders of magnitude.

As previously studied in a model of the yeast pheromone MAPK network, machine-based signaling does not exhibit the experimentally verified combinatorial inhibition [1]; instead, there is an upper limit on R_{max} that is realized near stoichiometric copy numbers and does not decrease as the concentration of scaffold increases (Fig. 3.5A, right). The ensemble models, however, do exhibit combinatorial inhibition, with peak R_{max} near stoichiometric scaffold concentration that

drops sharply at higher scaffold concentration (Fig. 3.5A, left). This decrease becomes more pronounced as the cascade depth becomes larger: as the valency of the scaffold increases, so does the combinatorial complexity of the system, and thus the influence of combinatorial inhibition is larger.

Varying scaffold numbers in the machine and ensemble signaling paradigm has little noticeable effect on either the remaining fitted Hill parameters, or the noise in response. Both cascade types maintain signal sensitivity (*i.e.* S_{50} decreases) upon raising the scaffold copy number. Machine-based models reach a limit upon which additional scaffold proteins make minimal difference, whereas the ensemble models' limit occurs due to a lack of response; as with R_{max}, combinatorial inhibition prevents signal throughput, and thus sensitivity to signal is an irrelevant quantity (Section B.2.2). Changes in response ultrasensitivity based on scaffold concentration are negligible, with minor decreases at high scaffold number (Section B.2.2). We also observe that a stoichiometric ratio of scaffolds to kinases exhibits a minimum in response variability (though the increase is less than an order of magnitude in all directions), whereas the machine model exhibits minor fluctuations in response variability with increased scaffold concentration but no global trend over the explored parameter space (Section B.2.2).

Another signaling property that shows a clear scaffold-dependent trend is the variation of T_{50} with scaffold concentration. In the machine models, increased scaffold numbers universally decrease the measured T_{50} over the explored range of cascade depths until reaching a limit between $5000 - 10000$ scaffold proteins (Fig. 3.5B, right). This occurs because higher scaffold concentrations raise the probability of initiating machine assembly, increasing the frequency of association between the scaffold and the first kinase in the cascade. This phenomenon is also recapitulated in the 2-kinase ensemble model, possibly due to the fact that the model essentially builds a 2-subunit signaling machine (though the reduced R_{max} resulting from combinatorial inhibition may also contribute to lower T_{50} values; Fig. 3.5B, left). For deeper ensemble cascades, increasing the scaffold copy number raises the T_{50}. This also results from combinatorial inhibition; as scaffold numbers grow, the time it takes to propagate signal also grows due to sequestration of signaling components

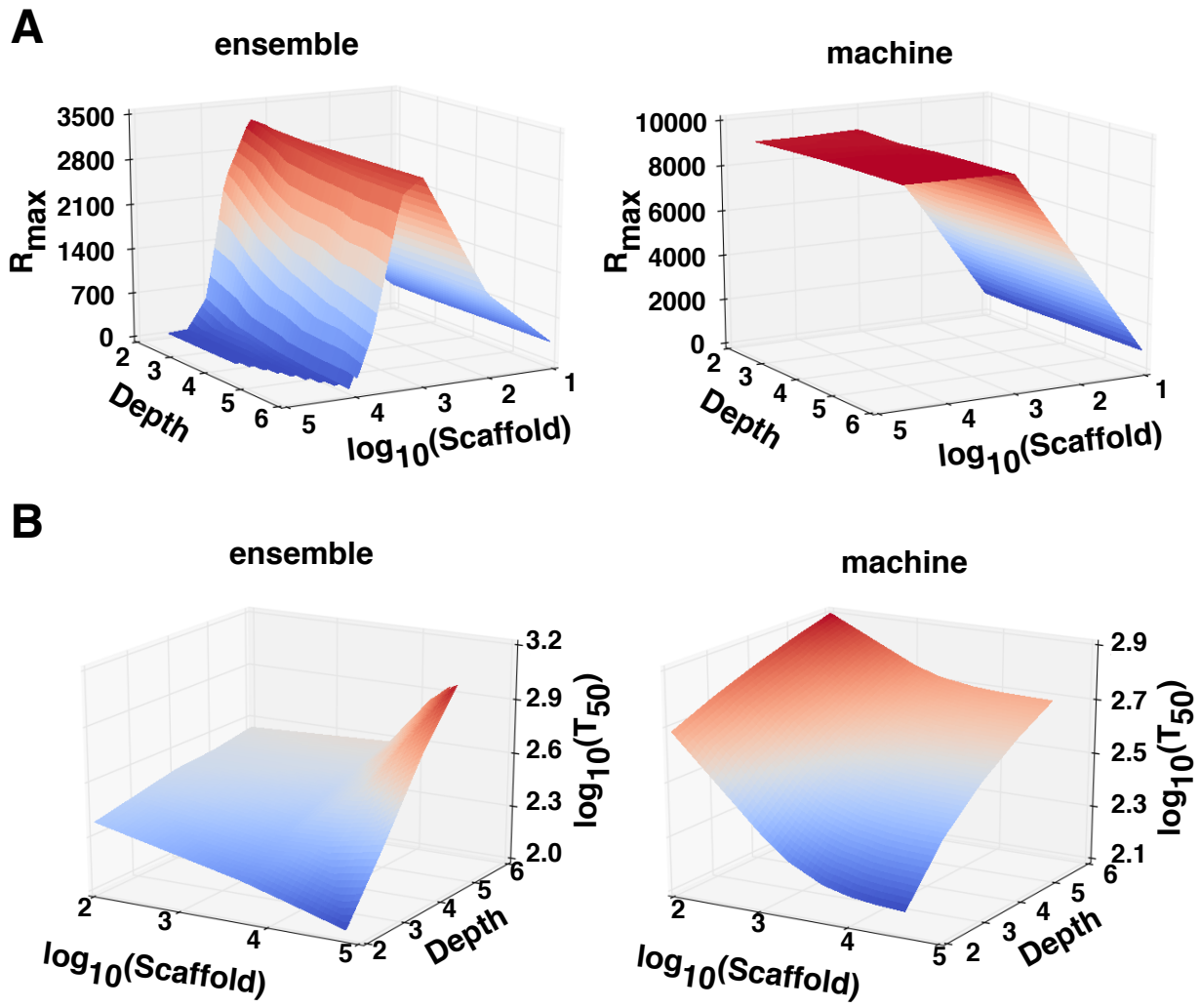


Figure 3.5: Scaffold concentration modulates select signaling behaviors. (A) Maximum response as a function of depth and scaffold number. Consistent with findings from prior experimental and theoretical studies [20, 21], we observe combinatorial inhibition due to high concentrations of scaffold proteins in the ensemble model simulations (left). Contrary to this, the machine model produces no such inhibitory effect since the hierarchical nature of signaling machine assembly prevents the combinatorial explosion of scaffold-based species that is present in the ensemble model [1]. (B) Response time as a function of depth and scaffold number at S_{max}. We observe a universal decrease in response time with respect to scaffold number in the machine model for all cascade depths (right), though in the parameter space considered here, this increase is less than an order of magnitude. Increasing scaffold numbers in the ensemble model, while showing faster responses in the 2-kinase cascade (likely due to the relative similarity between the 2-kinase ensemble model and the 2-kinase machine model rule structures, in addition to the reduced response due to combinatorial inhibition), displays slower response times for deeper cascades as a result of increased combinatorial complexity.

on different scaffold molecules.

3.2.5 Crosstalk

The ubiquity of crosstalk between signaling pathways (defined as one pathway's signaling components influencing another pathway's activity) in eukaryotic organisms is uncontested [90], and it is currently unclear how any degree of specificity is maintained in the face of this abundant crosstalk [34, 91]. One supposition is that scaffold proteins act as some sort of intracellular circuit board, directing signal transduction towards specific outputs for any given input [22]. It is likely that the assembly paradigm will influence the efficacy of crosstalk prevention in signaling cascades: in the absence of some sort of well-defined signaling complex (*i.e.* machine), it is unclear how scaffolds could prevent cross-pathway activation. To examine this, we adapted our three model types to include two pathways, each with a scaffold (except the solution model) and a shared kinase, in this case, the second in a 4-kinase cascade, K_2 (Fig. 3.6A).

Initially, we examined the effects for a scenario with maximum signal activating pathway A ($S_A = 10^5$ in the ensemble and solution models and $S_A = 10^2$ in the machine models), combined with a minimum signal activating pathway B ($S_B = 10^{-5}$ in the ensemble and solution models and $S_B = 10^{-8}$ in the machine models). We found that the solution model exhibited equal response from both pathways, despite stimulating only one (Fig. 3.6B, bottom). This is intuitive when considering that the third kinase, K_3, in each pathway competes equally for the pool of active shared K_2. Similarly, an active K_2 in the ensemble model may bind either scaffold A or scaffold B if it dissociates from pathway A's scaffold. Despite this fact, pathway A maintains a higher steady-state response than pathway B in this scenario (Fig. 3.6B, top). This is likely due to the additional biochemical events necessary for K_2 to activate pathway B. Upon its activation, the shared kinase can immediately phosphorylate the third kinase in pathway A, assuming $K_{3,A}$ is already present on pathway A's scaffold protein. Activation of $K_{3,B}$ requires an additional dissociation event (as mentioned above) and an association event (K_2 binding pathway B's scaffold), presumably resulting in a lower absolute response. This can be visualized via causality analysis tools present in the KaSim

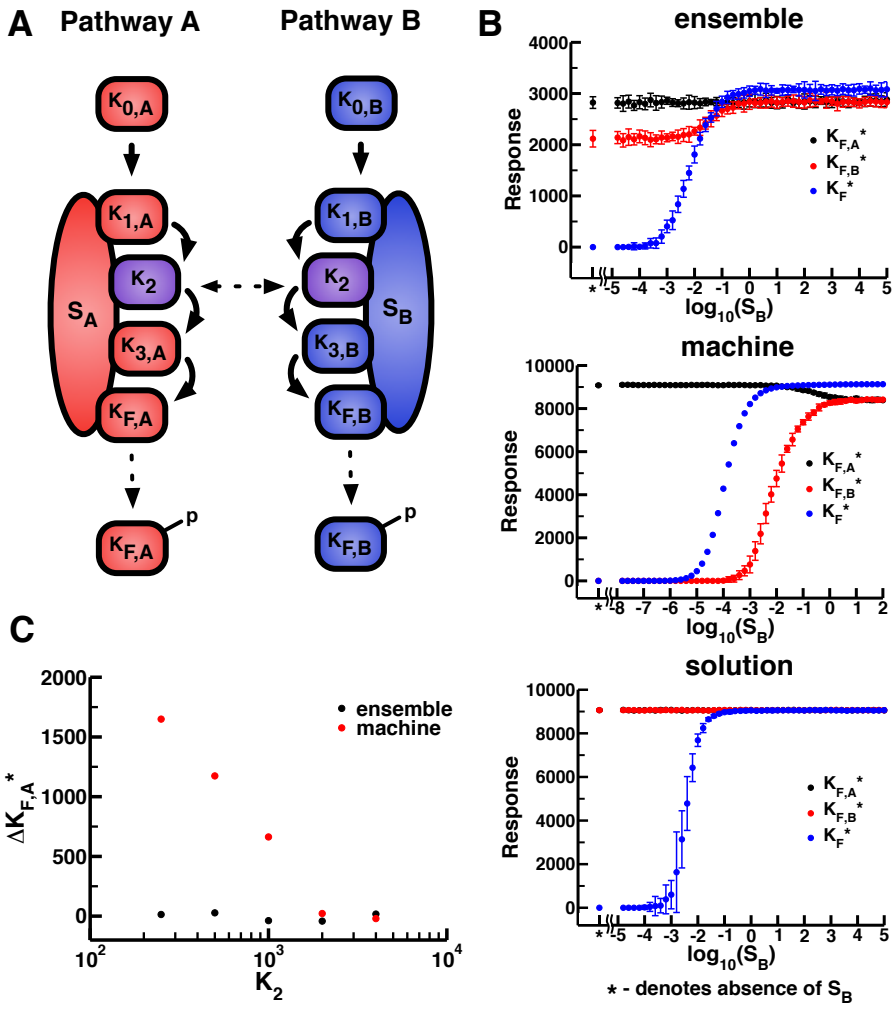


Figure 3.6: Crosstalk in the three signaling paradigms. (A) Schematic of crosstalk in scaffold-based signaling networks. In this figure, red components belong exclusively to pathway A while blue components belong to pathway B. The second kinase in both cascades (purple) is shared. Here, solid lines represent activation events, while dotted lines show translocation. In the solution model, kinases bind to one another, thus both $K_{1,A}$ and $K_{1,B}$ are capable of activating and binding K_2, which then binds and activates both $K_{3,A}$ and $K_{3,B}$. (B) Various cascade outputs (y-axis) as a function of log-transformed signals in pathway B (x-axis); pathway A is exposed to maximum signal ($S_A = 10^5$ for ensemble/solution simulations and $S_A = 10^2$ for machine simulations) for all data points. Black and red points indicate pathway A and B response, respectively. As a reference, blue points show the response for a single pathway model stimulated with S_B-strength signal. (C) Difference in pathway A response (y-axis) between a model with maximal stimulation of pathway A and minimal stimulation of pathway B and a model with maximal stimulation of both pathways as a function of the number of shared kinases (K_2; x-axis). As seen in panel (B), maximal activation of both pathways in the machine signaling paradigm introduces a decrease in output relative to maximal activation of only pathway A, whereas this difference is negligible in the ensemble model. This occurs when K_2 is the limiting factor in the signaling cascade (*i.e.* the component with the lowest copy number).

software package (see Section B.5) [15, 92]. Finally, no inappropriate cross-pathway activation exists in the machine model in this scenario (Fig. 3.6B, middle) since assembly of the signaling machine requires that pathway B's first kinase is active, which occurs very infrequently due to the low external activation of pathway B. In both the ensemble and machine models, increasing the signal input to pathway B eventually causes it to respond at levels similar to those of pathway A (Fig. 3.6B).

The machine model is thus the only signaling paradigm we tested that prevents one pathway from activating a second pathway where the second has no (or minimal) signaling input. However, an alternative form of crosstalk still arises in the machine paradigm and can be observed in Fig. 3.6B. In the case of our initial crosstalk models, the shared kinase is the limiting factor in signal transduction (*i.e.* the component with the smallest copy number) since its per-pathway concentration is halved. Competition for this kinase alters signal throughput, albeit in a different way than inappropriate activation of another pathway's output. In this case, the activity of one pathway is reduced when its components are recruited to another active pathway, and the output of pathway A actually *decreases* as pathway B becomes fully active (Fig. 3.6B, middle). To better characterize this phenomenon, we calculated the difference in $K^*_{F,A}$ between two cases: case 1, where pathway A is maximally active and pathway B is inactive, and case 2, where both pathways are maximally active. We represent the difference between case 1 and 2 as $\Delta K^*_{F,A}$. Due to the sequestration of the shared kinase, the machine model with a K_2 concentration of 500 molecules (*i.e.* half the concentration of the scaffold) has a $\Delta K^*_{F,A} > 1000$, indicating that full activation of pathway B can reduce the total pool of active $K_{F,A}$ by 10% (Fig. 3.6C). As one would expect, this drop in response output is mitigated with an increase in K_2 concentration: doubling the K_2 concentration relative to the scaffold (*i.e.* $K_2 = 2000$) results in essentially identical response from the first pathway regardless of the second pathway's level of stimulation. Thus, even if separate scaffolds nucleate formation of a machine-like signaling complex in two pathways, establishing true independence between those pathways requires detailed knowledge of the relative concentrations of the scaffold and any kinase that is shared between them. Interestingly, limiting K_2 concentrations do not generate similar

behaviors in the ensemble models (Fig. 3.6C), likely due to the fact that the shared kinase is not sequestered in an assembled (or assembling) signaling complex.

3.3 Discussion

Our results clearly indicate that the dynamic features of a signaling cascade can be drastically influenced not just by the presence of a scaffold protein, but also how the kinases in that cascade assemble onto the scaffold itself. These findings are summarized in Fig. 3.7. Strikingly, we found that only two of the response parameters we considered were similar between the ensemble and machine signaling paradigms. The most notable of these was the fact that presence of a scaffold protein in the cascade universally reduces noise and variability in molecular responses. Suppression of noise would clearly be advantageous in cases where individual cells must gather accurate information about their environment, such as determining if a potential mating partner is present [1, 78, 93]. This is likely related to the fact that scaffolds also generally linearize dose-response behavior, preventing the massive increase in ultrasensitivity that generally occurs as kinase cascades become deeper [86, 87].

The majority of the dynamic features we considered, however, showed strong dependence on how the kinases actually assemble onto the scaffold itself (Fig. 3.7). Machine-like structures generate higher absolute levels of output than ensembles, but require much longer times to achieve those responses, especially when signals are near the half-maximal level. Ensembles, on the other hand, can exhibit high degrees of combinatorial inhibition if scaffold concentrations are not tightly maintained near stoichiometric concentrations. The two assembly paradigms also have very distinct behaviors in terms of how they influence crosstalk. The machine model exhibits complete insulation from inappropriate activation by other pathways with shared downstream components, while the ensemble model does not. However, our models predict that scaffold proteins will reduce cross-pathway activation even in the ensemble case, improving signaling specificity relative to cascades that have no scaffold at all.

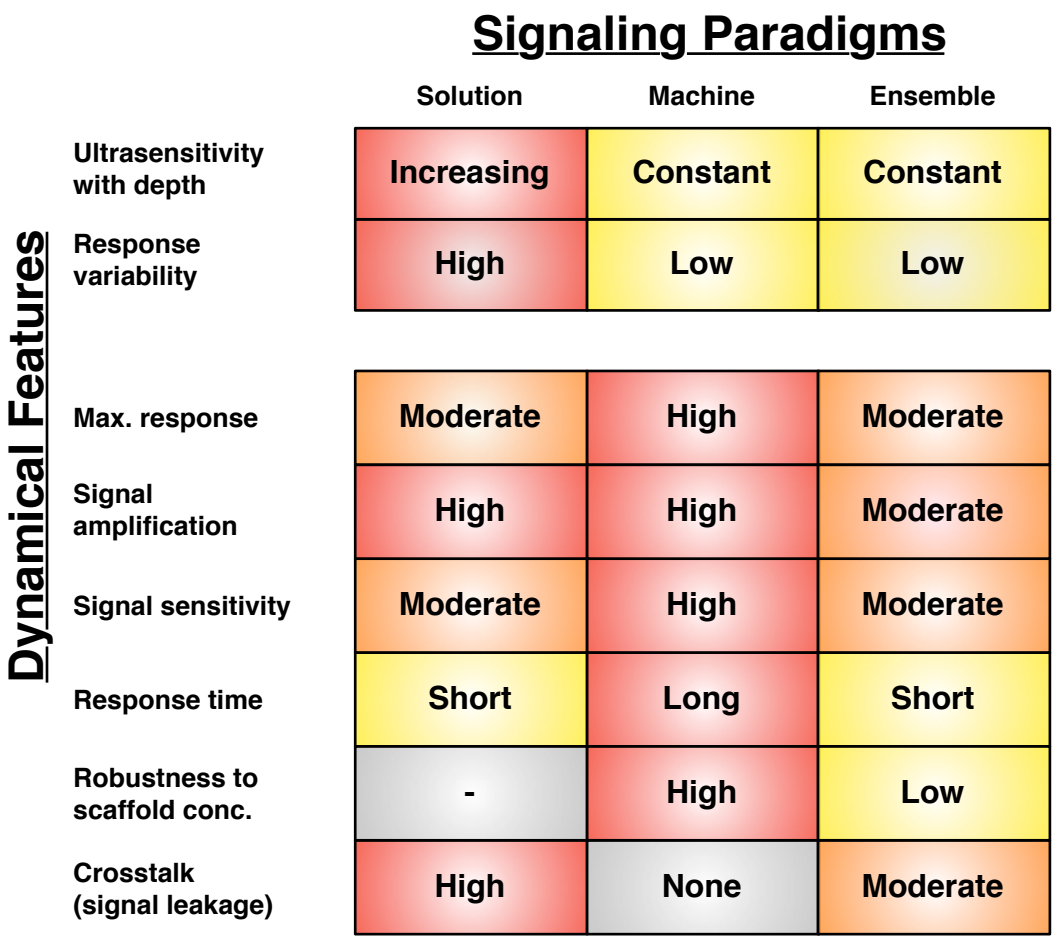


Figure 3.7: Comparison of various features between signaling paradigms. The three columns represent the three distinct types of signaling models considered in this work. Each row corresponds to a different dynamical feature. For two of these features, namely the variability of the response and the change in ultrasensitivity as cascades become deeper, the two scaffold-based signaling paradigms demonstrate similar behavior. In all other cases, however, the manner in which scaffold-based signaling complexes assemble is as important as whether or not the cascade uses a scaffold in the first place. Note that the results summarized in this figure are for unsaturated models of varying depths.

In general, these observations suggest that various assembly paradigms could play strikingly different evolutionary roles [22]. The nature of the signaling machine's complex structure and hierarchical assembly is reminiscent of highly-conserved multi-subunit proteins like the ribosome [30]. In this paradigm, the lack of combinatorial inhibition and decreased signaling time with increases in scaffold concentration indicate a resistance to fluctuations in protein concentration, which might arise due to the inherent noise in gene expression or from other, possibly "extrinsic," sources [5]. These traits couple well with scaffold-specific, but paradigm-independent, properties, such as reduced dose-response ultrasensitivity and noise in response (Figs. 3.3, 3.4). Scaffold complexes that assemble like machines can thus provide finely tuned and phenotypically robust behaviors. However, this type of multi-subunit protein might be difficult to evolve in comparison to the ensemble paradigm, since the scaffold would need to evolve extensive allosteric communication among its subunits in order to enforce hierarchical assembly (*e.g.*, the fact that kinase i will not bind the scaffold until kinase $i-1$ is already present in the complex, Fig. 3.1). Adding a new kinase to the cascade, or generating an entire signaling machine *de novo*, would thus likely require a rather lengthy process of evolving those constraints. In contrast, adding a new kinase to the ensemble model simply involves adding the relevant binding domain somewhere in the scaffold. Interestingly, extensive experimental work has shown that Ste5, the prototypical MAPK scaffold, can easily accommodate this kind of novel interaction, often generating highly functional dynamics just by adding new interactions or shuffling existing ones [13, 76, 78]. Ensembles thus exhibit a much higher degree of functional plasticity, generating weak regulatory linkage among signaling components and enabling the rapid evolution of new phenotypes [25]. Weak linkage, coupled with other ensemble-specific features (*e.g.* noise suppression, fast responses to signal) could provide strong fitness advantages in rapidly changing environments. In essence, scaffold proteins in the ensemble paradigm facilitate the evolution of additional cellular functionality (*e.g.* "rewiring" signaling pathways) whereas scaffold proteins in the machine paradigm better conserve existing cellular functions (*e.g.* reliably constructing ribosomes) [25].

At a more fundamental level, this sort of work can drive hypothesis development as well as

inform experimental design. For the former suggestion, we have shown that straightforward theoretical work can confirm or deny the feasibility of seemingly intuitive hypotheses. In this case, the relative technical simplicity of rule-based modeling both revealed and facilitated the characterization of the potential for signal amplification in scaffold-dependent signaling networks, where scaffolding was previously thought to actively prevent amplification [22, 23]. Formal methods and theory can also serve as a companion to synthetic biology. Given a sufficiently well-understood interaction network, mechanistic models could elucidate non-intuitive dynamics that might result from the introduction of novel proteins or domain recombinations [13]. Ultimately, we expect that this sort of theoretical investigation (either independent or coupled with experimental work) will become commonplace as a simple, inexpensive mechanism to better understand biological phenomena.

3.4 Methods

We performed stochastic simulations as well as causality analysis using KaSim and the Kappa rule-based modeling language, and we employed the BioNetGen software package for deterministic simulations [14–16]. Stochastic simulations were run until reaching an empirically determined steady-state or 10^5 seconds in simulation-time, due to the computationally intensive nature of exact agent-based Doob-Gillespie numerical simulations [89]. Both xmgrace (2D plots) and matplotlib (3D plots with linear interpolation) were used for data visualization, and custom analytical tools were developed in Python (available upon request).

Parameter	Values
Signal range $\left(\frac{V_{max,K_1}}{V_{max,P_1}}\right)$	10^{-5} to 10^{5} by $10^{1/5}$ (e, s) 10^{-8} to 10^{2} by $10^{1/5}$ (m)
Phosphatases (molecules, $i > 1$)	50 to 500 by 50, 750, 1000 (all)
Scaffolds (molecules)	10, 100, 1000, 2000, 5000, 10000, 20000, 40000 (all)
Kinases (molecules)	1000 (all)
K_M (molecules)	$\sim 10^5$ (unsaturated, $k_{on} = 1 \times 10^{-5}$ molecule^{-1} s^{-1}) ~ 1 (saturated, $k_{on} = 1$ molecule^{-1} s^{-1})

Table 3.1: Parameters used in our simulations. Note that all possible parameter combinations were not necessarily explored in this work. Abbreviations: Ensemble (e), Machine (m), Solution (s), Depth in cascade (i), Michaelis constant (K_M)

Chapter 4

The Noise is the Signal: Information Flow in Single Cells and Cellular Populations

4.1 Introduction

Signaling networks allow cells to sense intra- and extra-cellular concentrations of cytokines, nutrients, ions, *etc.*, and execute both discrete and continuous changes in cell state in response to those signals [1, 6, 86, 94, 95]. Apoptosis and commitment to cell division are typical of binary responses, whereas directed cell movement and induced gene expression are typical of continuously variable responses [1, 78, 93, 94]. Dysregulation of intracellular signaling has been implicated in a wide range of diseases including cancer, chronic inflammation, neurodegeneration, *etc*. [96], and developing a fundamental understanding of cellular information processing is instrumental in developing rational strategies aimed at treating those diseases [86, 97]. While signaling networks have been the subject of intense experimental and theoretical study for decades, it has only recently become possible to measure the response to signals at the level of individual cells [2, 6, 9]. These studies have revealed that signaling networks are subject to significant noise, which manifests itself within single cells as stochastic fluctuations in the activities of signaling proteins and as cell-to-cell variability within genetically identical cell populations [6, 7, 95, 97–103]. In some

cases, noise does not prevent individual cells from reliably making decisions. For example, yeast cells must decide whether or not to arrest the cell cycle in order to attempt mating when exposed to pheromone from other yeast cells, a critical cell-fate decision [104]. Work by Doncic, *et al.* has thoroughly characterized the mechanisms of this decision-making process, finding molecular markers (*i.e.* nuclear export of Whi5) that accurately predict a cell's decision to arrest or commit to division [105]. Interestingly, the decision to arrest the cell cycle implements signaling motifs that integrate information over multiple time scales (from rapid sensing of pheromone gradients on the order of seconds to the history of prior signaling events across multiple generations) allowing fine-grained control over this decision [106, 107]. In contrast, the presence of high levels of noise (and thus low accuracy in prediction of cellular response from a signal) in human signaling networks seems ubiquitous, and the ultimate physiological role of this nongenetic heterogeneity remains unclear given the evolution of reliable decision-making in other metazoan cells [105–107].

Information theory [26] provides a powerful analytical framework for quantifying the impact of noise on the ability of a system to transmit information. Levchenko and co-workers pioneered the application of information theory to signaling in mammalian cells [9] with the concentration of an extracellular ligand (*e.g.*, the inflammatory cytokine TNF-α) serving as the input to a (potentially noisy) intracellular signaling network (or channel), ultimately leading to a downstream response that can be experimentally measured (*e.g.* the nuclear translocation of NF-κB) . The information carried by the channel is quantified by the mutual information, I:

$$I(X;Y) = \int\int p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \mathrm{d}x\mathrm{d}y, \tag{4.1}$$

where X is the signal, Y is the response, $p(x,y)$ is their joint distribution and $p(x)$ and $p(y)$ are their corresponding marginal distributions [9, 26]. The base of the logarithm determines the units of the mutual information: the conventional base 2 quantifies information in "bits." This quantity can also be thought of as a measure of correlation between two random variables; with high mutual information, knowledge of one variable allows reliable prediction of the other [108].

Since the value of I depends on the input distribution, the mutual information of a signaling channel represents a combination of the properties of the signal and the intrinsic limits of the channel itself. It is thus problematic to use I as a measure for evaluating or comparing information flow in various signaling networks, since the *in vivo* distribution of signal values is rarely known. As a result, it is often more informative to focus on the maximum possible information that a channel can carry, or the channel capacity, C:

$$C = \sup_{p_X(x)} I(X;Y), \tag{4.2}$$

where the supremum (the least upper bound) is evaluated over all possible choices of the probability distribution of the input. The channel capacity is an inherent feature of the channel: the larger the value, the more information that a channel can theoretically transmit [9, 26].

In the case of cellular responses to TNF and other cytokines, Cheong *et al.* found that the channel capacity is generally less than 1 bit for molecular responses at the single-cell level (these values are summarized in Table 4.1, entries 1-4, 12) [9, 109–111]. The implication, then, is that many intracellular signaling networks cannot reliably distinguish between the presence or absence of TNF, EGF, and other signaling molecules ($C < 1$ bit, Table 4.1) [9]. More recent work has focused on characterizing various strategies that cells might employ to achieve higher levels of information transfer. For instance, Lee *et al.* demonstrated that mechanisms such as fold-change detection (in which cells are sensitive to the ratio between a steady-state and induced signal) decrease the impact of noise on the propagation of TNF-induced signals [28]. As described below, however, we found that the channel capacity between TNF concentration and the downstream transcriptional response remains below 1 bit despite the use of fold-change detection in this system (entry 5, Table 4.1). Wollman and co-workers recently demonstrated that using multiple time points from the trajectory of a molecular response (*e.g.* Erk activation over time) can significantly increase channel capacities. While this dynamic approach to information flow clearly can increase C, it is currently unclear how cells might actually implement this kind of mechanism at the molecular level [2].

One limitation of previous work is the focus on measuring the activity of signaling intermediates, such as nuclear localization of the NF-κB transcription factor, rather than a cellular phenotype

Signal (molecular)	Response (molecular)	C (bits)	Data	Calculation
1. TNF	NF-κB	0.92 ± 0.01	[9]	[9]
1.1. TNF	ATF-2	0.85 ± 0.02	[9]	[9]
1.2. TNF	NF-κB & ATF-2	1.05 ± 0.02	[9]	[9]
2. PDGF	NF-κB	0.67 ± 0.01	[9]	[9]
2.1. PDGF	ATF-2	0.74 ± 0.01	[9]	[9]
2.2. PDGF	NF-κB & ATF-2	0.81 ± 0.02	[9]	[9]
3. EGF	Erk (fold-change)	0.60 ± 0.03	[109]	[9]
4. UDP	Peak Ca^{2+}	1.22 ± 0.03	[110]	[9]
4.1. UDP	Integrated Ca^{2+}	1.07 ± 0.02	[110]	[9]
5. TNF	A20 transcripts	0.84 ± 0.09	[28]	this work
6. TRAIL	Casp-8 activity	1.01 ± 0.01	this work	this work
7. TRAIL	Casp-8 activity (live cells)	1.01 ± 0.01	this work	this work
8. TRAIL	Casp-3 activity	0.56 ± 0.01	this work	this work
9. Casp-8 activity	Casp-3 activity	1.23 ± 0.01*	this work	this work
10. Casp-8 activity	cell decision	0.51 ± 0.01*	this work	this work
11. α-factor	*pFUS1*-GFP	2.26 ± 0.04	[78]	this work

Signal (position)	Response (molecular)	C (bits)	Data	Calculation
12. Embryo perimeter	Phosphorylated Erk	1.61 ± 0.05	[111]	[9]

Signal (position)	Response (motion)	C (bits)	Data	Calculation
13. Bacterium	neutrophil motion	1.82 ± 0.11		this work
14. cAMP	*Dictyostelium* motion	2.19 ± 0.08	[112]	this work

Signal (molecular)	Response (population)	C (bits)	Data	Calculation
15. TRAIL	% dead (HeLa; resampled)	2.44 ± 0.02	this work	this work
16. TRAIL	% dead (HeLa; FACS)	3.41 ± 0.02	this work	this work
17. TRAIL	% dead (MCF10A)	3.38 ± 0.02	this work	this work

Table 4.1: The channel capacity for population-level response in HeLa cells using FACS was calculated using 1000 cells per TRAIL concentration and all population-level channel capacities were calculated using 100 independent populations. Confidence intervals indicate 95% confidence level in estimator. Asterisk indicates bin incrementation was capped for computational efficiency (*i.e.* the value is likely an underestimate).

as an output [9, 28]. This makes it difficult to interpret the functional significance of low channel capacities. We therefore focused our analysis on an unambiguous terminal phenotype: life or death as regulated by TNF-Related Apoptosis-Inducing Ligand (TRAIL). TRAIL induces apoptosis by binding to cell surface receptors, initiating formation of death-inducing signaling complexes or DISCs. These complexes then activate initiator caspases (ICs), starting a sequence of biochemical events resulting in mitochondrial outer membrane permeabilization (MOMP). The release of numerous mitochondrial proteins into the cytosol and subsequent formation of the apoptosome then promotes activation of the effector caspases (ECs), ultimately leading to cell death (Figure 4.1A). Dramatic cell-to-cell variability has been observed in the responses of clonal cell cultures to TRAIL (and other death ligands): whereas a subset of cells dies within 2-8 hr of ligand exposure, others survive indefinitely. When these survivors are re-assayed for TRAIL sensitivity following outgrowth, the same fractional killing is observed, showing that variability is a stable property of the cell population. Molecular studies have shown that this variability arises from extrinsic noise in receptor-to-caspase signaling networks [6, 98, 99].

We found that the channel capacity between TRAIL dose and IC or EC activity (measured at the single-cell level) is significantly less than 1 bit, similar to other molecular responses to cytokine signals (Table 4.1, entries 6-8). Interestingly, however, we found that the channel capacity between TRAIL dose and the *fraction of cells that die* at that dose is much higher, around 3-4 bits, indicating that TRAIL-induced apoptosis may have evolved to control cellular populations rather than individual cells. This process is known as gain adaptation, meaning that this network has evolved to maximize information transmission to cellular populations by aligning the dose-response curve of the signaling network to the natural probability distribution of TRAIL [114, 115]. We thus developed a simple mathematical model that allowed us to characterize the fundamental trade-off between the amount of information individual cells can have about their environment and the amount of information that can be used to control decision-making at the population level. We also demonstrated that the low channel capacities generally observed for single-cell responses are not a result of inherent biophysical limitations: by analyzing data on eukaryotic chemotaxis and

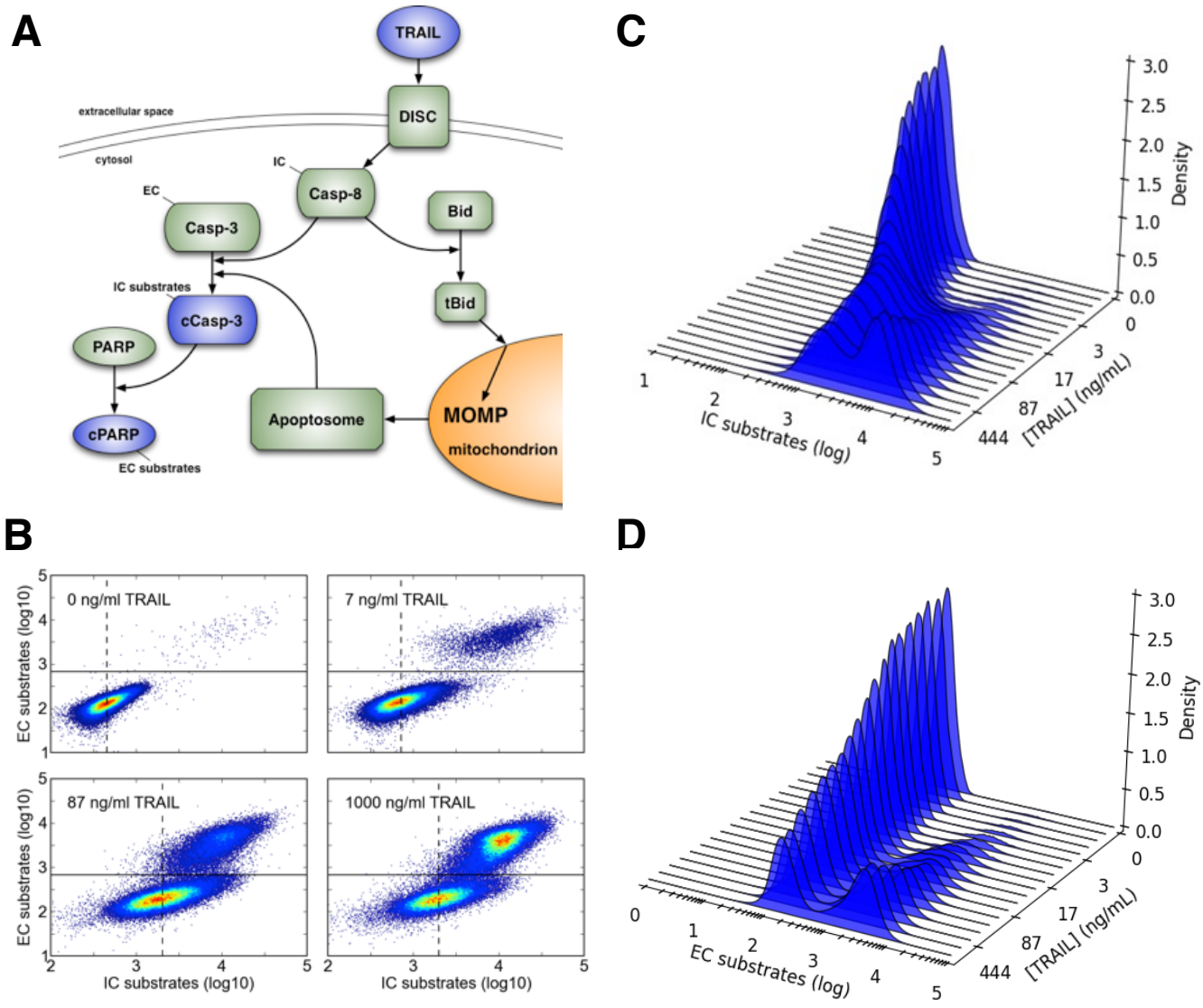


Figure 4.1: (A) TRAIL activates the extrinsic apoptosis signaling pathway through activation of the initiator caspase (IC) Casp-8 via death-inducing signaling complexes (DISCs). Active Casp-8 then activates the effector caspase (EC) Casp-3, via two mechanisms: direct cleavage and mitochondrial outer membrane permeabilization or MOMP, which induces formation of the apoptosome, another activator of Casp-3 [99]. (B) Measurement of cleaved EC and IC substrates by flow cytometry [99] show that HeLa cells have a highly variable response to TRAIL across a wide range of doses ($n = 60,000$ cells per TRAIL dose). The solid line is the minimum density in the bimodal EC response ($\sim 10^{2.8}$) and acts as a threshold for apoptosis, whereas the dashed line marks the average IC response for non-apoptotic cells. (C & D) We used kernel density estimators in the R statistical software package [113] to estimate TRAIL-dependent response distributions for IC (C) and EC (D) activity. The fraction of EC activity above the threshold is proportional to the number of apoptotic cells [99] indicating that approximately 50% of cells survive the maximum TRAIL dose.

mating for yeast cells, we found that some signaling networks are capable of transmitting well over 2 bits of information, allowing more precise control of individual cells' behavior than simple binary decision-making. Ultimately, our work suggests that noise in cell signaling is likely highly regulated. When the key physiological output is the behavior of a single cell (as in chemotaxis or mating), noise is likely suppressed [2, 28] to enable high levels of information transfer to those cells. When the key physiological output is the fraction of cells in a tissue or population that undertake a certain decision (*e.g.* commitment to apoptosis or cell division), noise is likely exploited so that information can be transferred at the population level.

4.2 Results

4.2.1 Individual cells responding to TRAIL exhibit low channel capacity

To measure the channel capacity of the extrinsic apoptosis signaling cascade (Figure 4.1A), HeLa cells were treated with TRAIL for 11 hr over a range of ligand concentrations from sub- to super-physiological, and molecular responses in single cells were measured by flow cytometry (13). The level of cleaved caspase-3 (cC3) served as a measure of the time-integrated activity of receptor-proximal ICs and cleaved PARP (cPARP) served as a measure of downstream EC activity (Figure 4.1A). Previous studies have shown that TRAIL exposure results in a dose-dependent increase in IC activity that varies significantly from cell-to-cell; in any single cell, when IC activity exceeds a threshold set by anti-apoptotic Bcl-2 proteins, ECs are activated and the cell proceeds inexorably to death (Figure 4.1B-D) [6, 99].

While Eqs. 4.1 & 4.2 seem concise at first glance, estimation of the mutual information and channel capacity is a nontrivial challenge, and numerous approaches have been proposed and implemented [116–118]. In order to facilitate comparison between our calculations and those performed by Cheong *et al.*, we designed a software package to estimate mutual information based on the binning procedure they applied in their work (see Section C.1.1) [9, 119]. This software is freely available as an open-source project (`https://github.com/ryants/EstCC`). Using this

software and the distributions of IC and EC activity in single cells, we calculated a channel capacity between TRAIL dose and IC activity of $C \approx 1.01$ bits and between TRAIL and EC activity of $C \approx 0.5$ bits (entries 6 and 8, Table 4.1), with the latter being far lower than the 1 bit of information needed to reliably make the decision to undergo apoptosis based on extracellular TRAIL levels. We observed similar values for TRAIL to IC channel capacity in surviving cells (*i.e.* those with low cPARP levels, see Section C.2.3). As an additional control, we calculated the channel capacity between IC activity and both EC activity and cell fate obtaining a $C \approx 1.23$ and 0.51 bits, respectively (entries 9 and 10, Table 4.1). This confirms that measured IC activity is a relevant intermediate signal for extrinsic apoptosis since it contains similar levels of information regarding the binary cell-fate decision than TRAIL, and that noise accumulates in the upstream stages of apoptosis signaling pathway [6]. High levels of noise and low levels of information transfer are thus a feature of the TRAIL network across multiple biologically relevant measures.

4.2.2 Population response to TRAIL exhibits high channel capacity

However, when we examined the channel capacity between TRAIL dose and phenotypic response at the population level we obtained a very different result. The combination of noise and a threshold can allow a fraction of cells in a population to make a discrete decision in response to a signal [6, 7, 100–102, 120]. For either of two cell types (transformed HeLa and non-transformed MCF10a cells), we found that the fraction of cells surviving exposure to TRAIL gradually decreased as the concentration of ligand increased over a 103-fold range (Figure 4.2). The fraction of cells dying at any given TRAIL dose in both experiments showed comparatively little variance between replicate experiments (Figure 4.2). As a result of this relatively low variability, the channel capacity between TRAIL dose and the fraction of cells undergoing apoptosis was much higher than what we observed for the molecular response in single cells, between 3.4 and 4 bits depending on the population size (entries 15 and 16 in Table 4.1, and Figure 4.3). In essence, this indicates that cellular populations can respond in a meaningful way to a wider range of signal values and it is plain to see in the dose-response curves; where there are strongly overlapping response distribu-

tions over most of signal space for individual cells (Figure 4.1C & D), the population response distributions are more distinguishable from one another (Figure 4.2. Since the variability between dose-response replicates is at least partly technical in nature (*e.g.*, due to handling of the cells during the experiment), these values represent lower bound estimates of the true biological channel capacity.

4.2.3 Understanding the trade off between single-cell and population-level information transfer

To better understand how single-cell noise contributes to high channel capacity at a population level [7, 95, 100–102], we created an idealized model for intracellular signaling in which signal (S) and response (R) are related by a Hill function modified to account for noise:

$$R = (R_{\max} - R_{\min}) \cdot \frac{S^n}{S^n + K^n} + R_{\min} + \varepsilon \tag{4.3}$$

where K is the concentration of an input ligand that results in a half-maximal response, n is the Hill exponent (a measure of dose-response ultrasensitivity), $R_{\min}$ and $R_{\max}$ represent the range of average responses, and ε is a noise term sampled from a Gaussian distribution with mean $\mu = 0$ and variable standard deviation, σ [99]. It is important to note that this model is designed to be a phenomenological model and is not meant to precisely recapitulate experimental data. Estimation of the single-cell channel capacity in this model involves generation of a dose-response data set with N independent cells per M distinct concentrations of input signal, resulting in $N \times M$ ordered (S, R) pairs. By varying σ, we created dose-response data for specified levels of noise (Figure 4.3A) [99]. To simulate responses at the population level (*e.g.* the fraction dead), one must map individual cell responses to a discrete phenotype. Our model therefore assumes that individual cells exhibit a phenotypic response (*e.g.* cell death) when R exceeds a threshold value, (analogous to the threshold set by anti-apoptotic Bcl-2 proteins in apoptosis) [99]. At any given signal value we thus have the distribution of R values in a simulated population, and also the fraction of that

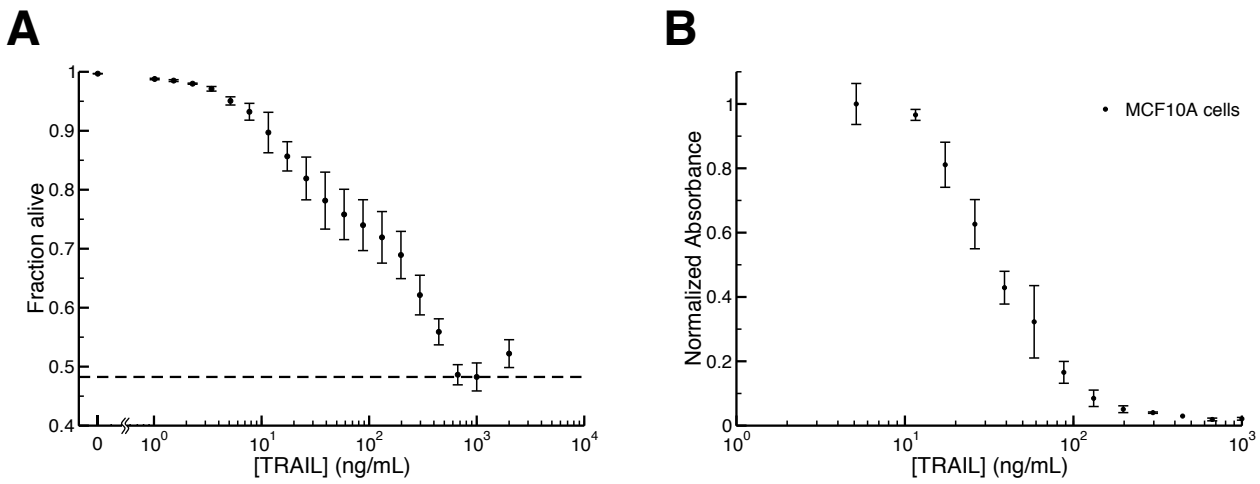


Figure 4.2: We used the threshold described in Fig. 4.1B to map data from HeLa cells to fractional survival at varying TRAIL doses. We recorded a maximal effect at [TRAIL] $= 1000\ ^{\mathrm{ng}}/_{\mathrm{mL}}$ (indicated by the dashed line); higher doses of TRAIL lead to less fractional killing in a "ligand squelching" effect that we have consistently observed for this system. Since the channel capacity represents a supremum over all possible probability distributions of input signals, we removed the final point ([TRAIL] $= 2000\ ^{\mathrm{ng}}/_{\mathrm{mL}}$) from our analysis without loss of generality. Error bars indicate sample standard deviation across 3 replicates of 20,000 cells each. (B) Fraction of MCF10A cells surviving TRAIL treatment as assayed by methylene blue staining [98] show a graded response similar to that of the HeLa population in (A). Calculation of the channel capacity based on this data yielded $C \approx 3.4$ bits, similar to the value for the HeLa cells (Table 4.1).

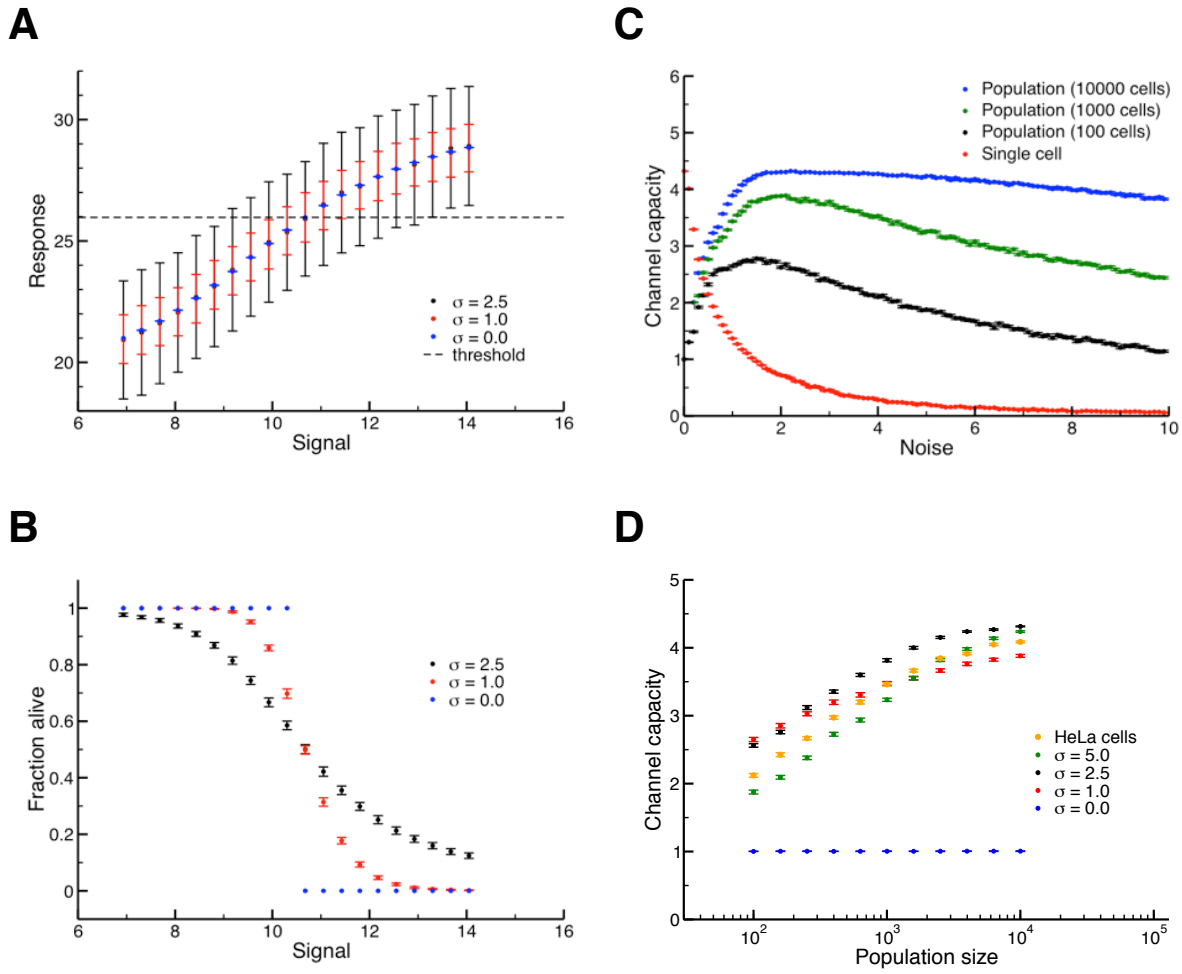


Figure 4.3: (A) Single-cell dose-response behavior in the initial model described in Eq. 4.3. The mean response and sample standard deviation of 1000 independent simulated "cells" is shown for various noise values, relative to a cell death threshold (dashed line). (B) Population dose-response behavior from $P = 100$ independent populations with $N = 1000$ cells per signal each. Individual cells' response map to either death or survival according to the threshold in (A); points correspond to the mean and sample standard deviation of the fraction of surviving cells. (C) A trade-off exists between single-cell and population-level channel capacity. Increasing noise decreases information transmission in single cells and simultaneously increases the population-level channel capacity up to an optimal noise value. (D) Channel capacity in our model correlates with population size in the presence of noise (green, black, red, blue). Additionally, the level of noise needed to maximize channel capacity changes as the population size grows, from low values ($\sigma \approx 1$ for $N = 100$) to higher values ($\sigma \approx 2.5 - 5$ for $N = 10,000$). Experimental population-level channel capacities (orange) were calculated by taking 100 random subsamples from the set of 60,000 HeLa cells for a range of population sizes.

simulated population that exhibit the phenotypic response.

For simulated populations of $N = 10^2$ to 10^4 cells, the model revealed a striking trade-off between the channel capacity for single cells and for cell populations. When noise is low, the response of individual cells is essentially deterministic, corresponding to a step-like change in the fraction of cells that die as S increases (Figure 4.3A & B, blue). At higher levels of noise, the response of individual cells obviously becomes more variable, but this corresponds to a gradual decrease in the fraction of "surviving" cells (Figure 4.3A & B, black and red) [6, 7, 100–102, 120]. As observed in the experimental data (Figure 4.2), the fraction of cells that respond at any given value of S displays relatively low variance between populations of the same size (corresponding to relatively small error bars on the red and black curves in Figure 4.3B).

Low noise thus leads to high channel capacities between S and R measured at the single-cell level (over 5 bits), but very low channel capacities between S and the fraction of cells that die (~ 1 bit, Figure 4.3C). As the level of noise increases, channel capacity at the single-cell level drops rapidly but at the population level it rises significantly, before falling again (Figure 4.3C). The level of noise that optimizes population-level channel capacity varies with the number of cells. For example, with $N = 10^2$ a maximum of $C \approx 2.75$ bits is achieved with $\sigma \approx 1$; with 10^4 cells $C \approx 4$ with $\sigma \approx 2.5$ (Figure 4.3D). It should be noted that the precise position of this maximum depends on the spacing of the signal values S: since lower values of σ correspond to a narrower population-level dose response (Figure 4.3B), re-sampling S values more finely within the transition region tends to decrease the value of noise that maximizes C (see the Section C.4.5). However, so long as there is some minimum spacing between the discrete S values to which a population can be exposed, or any error in generating precise values of S (*e.g.*, stochasticity in the production of cytokines by other cells), the maximum observed in Figure 4.3D occurs at standard deviations significantly larger than 0.

To examine the effect of population size in our experimental data, we randomly sampled sub-populations of HeLa cells from the total of 60,000 per TRAIL dose that we measured. This revealed a similar dependence of population-level information transfer on population size in our

experimental data (Figure 4.3D, orange). Taken together, our work demonstrates that the combination of a noisy signaling network with concomitantly low information transfer (Figure 4.1 and Table 4.1) and a threshold in initiator caspase activity [99] leads to robust information transfer at the level of cell populations (Figures 4.2 and 4.3).

4.2.4 Low channel capacities observed previously likely do not represent intrinsic biophysical limits

Although noise may ultimately support information transfer to cell populations, it is unclear if the phenomenon discussed above represents cells simply taking advantage of the existing noise in signaling systems, or if noise can be tuned up and down to favor fidelity in either single cells or population-level decisions. To explore this latter possibility, we considered two cases in which individual cells (rather than populations) must respond accurately to environmental stimuli. During *S.cerevisiae* mating, haploid **a** and α cells must determine if a suitable mating partner is sufficiently close for conjugation to be successful. Cells sense the local concentration of the mating pheromone α factor via a G protein-coupled cell surface receptor and a downstream MAP kinase signaling cascade; when a suitable partner is available for conjugation they reorient their cytoskeletons and initiate a complex transcriptional program [1, 78, 93]. As mentioned in the introduction, the decision to mate results in cell-cycle arrest [104], and so we would expect there would be an evolutionary pressure for individual yeast cells to have a relatively high level of information about the availability of mating partners in their environment. Even when considering only single-cell data from the pheromone-sensing network [78], and not the status of the cell cycle or prior history of signaling as did Doncic, *et al.* [105–107], we calculated $C \approx 2.26$ bits between α-factor dose and the transcriptional output, measured by a fluorescent reporter (entry 11, Table 4.1). This particular network thus demonstrates a notably higher level of fidelity than has been observed for molecular responses to cytokines in metazoan systems (Table 4.1).

Another example of a situation in which individual cells are the key biological actors is eukaryotic chemotaxis. We therefore analyzed a classic movie of a human neutrophil "hunting" a

bacterial cell, and a movie of a single *Dictyostelium* cell responding to cAMP emanating from a micropipette (both movies are available as supplemental files) [112]. Because migrating cells are polar, it is possible to define a cell-based coordinate system using standard tracking software (CellTrack) [121]. Like others working on distributions of directional movement [122], we defined the input as the angle between the chemoattractant (bacterium or micropipette) and the cell axis and the output as the angle of the cell's subsequent motion (Figure 4.4A). For both the neutrophil (a representative trajectory is shown in Figure 4.4B) and *Dictyostelium* we computed $C \approx 1.81$ and $C \approx 2.2$ bits (entries 13 and 14, Table 4.1), which is almost certainly a lower bound given that we are simplifying a 3D problem as a 2D search. From these data we conclude that signaling networks in single cells can encode more than 2 bits of information (possibly much more) demonstrating that previous observations of $C \approx 1$ are likely not due to the fact that the inherent noise in biochemical reaction networks limits channel capacities to below 1 bit.

4.3 Discussion

Our findings touch on two distinct and complementary aspects of information transfer in signal transduction: single-cell and population-level information processing. In the case of regulatory networks that control apoptosis, the key physiological variable is the fraction of cells responding at a given dose [101, 120]. In this case, low channel capacity at a single-cell level ($C < 1$) is a corollary of high capacity at a population level ($C \approx 3$ to 4). Said another way, achieving effective control over fractional responses requires a significant heterogeneity at the single-cell level [6, 7, 94, 100–102, 120], and thus low channel capacities when the responses of those cells are assayed at a single-cell level. Since many cytokines regulate population-level behaviors (*e.g.* control over neural progenitor cell proliferation and differentiation by EGF/NGF [94]), it is perhaps not surprising that channel capacities less than one bit have been observed in those cases (Table 4.1).

In contrasting cases where individual cells must precisely resolve signals to make decisions in a continuous response space (*e.g.* finding a mating partner, following gradients or hunting pathogens)

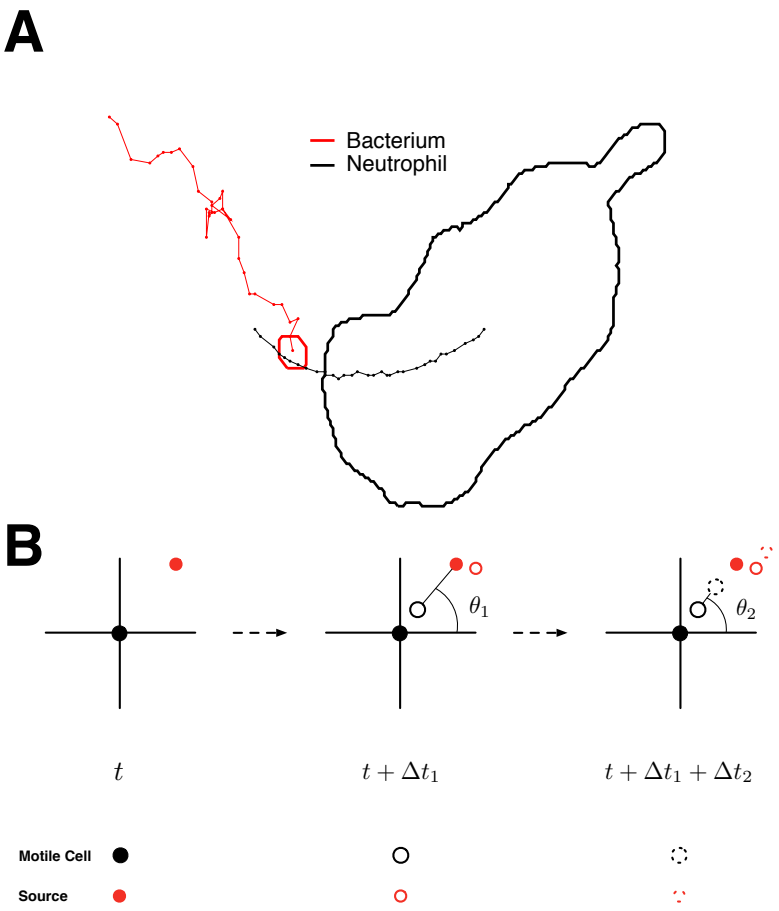


Figure 4.4: (A) A representative trajectory from the neutrophil movie. Points are centers of mass for the bacterium (signal source, red) and neutrophil (motile cell, black). The bold, outlined areas show the cells' perimeters in the trajectory's first frame. Clearly seen here is the bacterium's stochastic random walk-like motion and the neutrophil's smoother tracking of the resulting gradient. (B) Black circles represent the motile cell and red circles represent the signal source (either a micropipette or bacterium). Since cells do not instantaneously detect or respond to extracellular stimuli, filled, solid, and dashed circles represent the motile cell and signal source at initial (t), signal-delayed ($t + \Delta t_1$), and response-delayed ($t + \Delta t_1 + \Delta t_2$) time, respectively. We can then calculate the mutual information between the signal (θ_1) and response (θ_2) angles.

we find that the single-cell channel capacity is generally significantly higher than has hitherto been observed ($C \approx 2$, Table 4.1) (Figure 4.4). As mentioned above, we expect that this value is likely a lower bound, since our chemotaxis data is essentially a 2D projection of a process that often occurs in 3D space. Additionally, the data for the *Dictyostelium* calculation ($C \approx 2.19$, Table 4.1, entry 14) exhibits tight distributions around approximately 6 input angles, producing a maximum input entropy (and thus an upper limit on channel capacity) of slightly less than $\log_2(6) = 2.6$ bits. Since the estimated channel capacity is so close to this limit ($> 85\%$), we suspect that the calculated spatial channel capacity in this case would increase upon further sampling of signaling space (*i.e.* the relative angle between the source of the signal and the cell).

The clear conclusion from these findings is that low channel capacity at a single cell level ($C < 1$) does not reflect an inherent limit in the biochemistry of signal transduction, but rather a natural trade-off between the knowledge that individual cells have about their environment and the ability of multicellular organisms to control responses reliably at the population level. With respect to noise levels in these systems, two nonexclusive possibilities exist. The first is that networks that control cellular populations simply exploit noise that arises from stochastic fluctuations in transcription, protein synthesis and related processes whereas as chemotactic networks have evolved to suppress it. The second is that some signaling networks have actually evolved higher levels of noise than the underlying biophysics dictates [123–126]. In either case, the physiological importance of noise may explain why drugs that target cellular decision networks have difficulty eliciting complete population-level responses [97]. Understanding and ultimately exploiting biological noise is likely to be as important for therapy as it is for metazoan signaling.

4.4 Methods

4.4.1 Experimental methods

All experiments were performed by John Bachman at Harvard Medical School. HeLa cells were maintained in DMEM medium (Corning 10-013-CV) with 10% fetal bovine serum and 1% peni-

cillin/streptomycin solution (Life Technologies 15140-122). For TRAIL dose-response assays, HeLa cells were plated at a density of 250k cells/well in 12-well plates (Sigma SIAL0513), allowed to adhere overnight, and treated with varying doses of SuperKiller TRAIL (Axxora ALX-201-115) for 11 hours. Three replicate wells were used for each TRAIL dose to establish the technical variability of the assay. After treatment, medium containing dead cells was transferred to flow cytometry tubes (BD Falcon 352235) containing 2 ml FACS buffer (PBS + 10% fetal bovine serum); cells remaining in the wells were removed by trypsinization, added to the corresponding tubes, pelleted by centrifugation and fixed in 4% paraformaldehyde for 30 minutes. After fixation cells were washed twice in PBS and permeabilized in 100% methanol overnight at -20C. Cells were stained with primary antibodies to cleaved caspase 3 (rabbit anti-cleaved caspase 3, BD 559565) and cleaved PARP (mouse anti-cleaved PARP, BD 552596) 1:250 in FACS buffer (PBS + 0.1% Tween-20) for 1 hour at 25C. Cells were washed twice in PBS-T, then treated with secondary antibodies: Alexa-488 donkey anti-rabbit IgG (Life Technologies A-21206) and Alexa-594 donkey anti-mouse IgG (Life Technologies A21203), 1:500 in FACS buffer for 1 hr at 25C. Cells were washed in PBS-T, resuspended in PBS, and counted on a flow cytometer (BD LSRII), with 20,000 cells analyzed per experimental replicate.

MCF10A cells were obtained from J. Brugge (Harvard Medical School, Boston, MA) and cultured as described [127]. For TRAIL dose response assays, MCF10A cells were plated in 96-well plates (Corning 353072) and treated with varying doses of SuperKiller TRAIL for 11 hours. After treatment the cells were washed with PBS and the density of viable cells was assayed by methylene blue staining as described previously [98].

4.4.2 Estimating mutual information

The code used to calculate the channel capacity was based primarily on the description of mutual information estimation in Cheong *et al.*'s supplementary texts [9] but was modified in a few ways. Instead of calculating the average mutual information of the "plateau" region of bins, we take the maximum mutual information such that at least one mutual information estimate from 3 random-

ized data sets is not significantly different from 0 in its intercept estimate (95% confidence). The error bars are then the 95% confidence intervals for the estimate of the intercept (*i.e.* the mutual information extrapolated to infinite sample size).. We used this method for all mutual information calculations performed in this work in order to ensure accurate comparison between values. The source code can be found at `http://github.com/ryants/EstCC` and a complete description of the estimation procedure can be found in Section C.1.1.

4.4.3 Model construction

All calculations involving the model seen in Eq. 4.3 were performed with the following (arbitrarily chosen) parameters: $K = 10$, $n = 6$, $R_{\max} = 30$, and $R_{\min} = 20$. Our range of 20 signal values was chosen such that the minimum and maximum response values in our data set were 10% above and 10% below $R_{\min}$ and $R_{\max}$, respectively, and the remaining 18 values were evenly distributed in between. In this way, the Hill coefficient governing the slope (or ultrasensitivity) of the response, and the sampled signal space, minimally impacts the channel capacity calculation (see Section C.4). The threshold value was chosen so that half of the signal values produce an average response below the threshold and half produce an average response above the threshold. In the absence of noise, this selection would result in a channel capacity of 1 bit.

4.4.4 Spatial channel capacity calculation

From the 2D data provided by the CellTrack program, we calculated the mutual information between the initial angle created from the motile cell (θ_1) and the signal source and the resulting angle of motion of the motile cell (θ_2) as mentioned in the main text (Figure 4.4). Since information transmission does not occur instantaneously, we introduced two time-delay factors: Δt_1, which is the time necessary for the motile cell to detect the signal, and Δt_2, which is the time required for the neutrophil to respond to extracellular information. We calculated $C(\theta_1, \theta_2)$ for a range of $(\Delta t_1, \Delta t_2)$ pairs and reported the maximal value in Table 4.1.

Chapter 5

Intrinsic Limits of Information Transmission in Biochemical Signaling Motifs

5.1 Introduction

Signaling networks enable cells to sense information about their environment in order to adapt appropriately to changing conditions. Quantifying the reliability of communication has long been the domain of information theory [26], and information theoretic concepts have been relevant to certain many of biology for quite some time (most notably neuroscience and bioinformatics) [114–116, 119, 128–130], but have not been applied to systems biology until recently. These ideas and quantities are now becoming increasingly relevant for understanding information transmission via signal transduction networks, however, and to the corresponding cell-fate decisions that must be made on the basis of environmental cues [2, 9, 103, 131, 132]. In the context of cell signaling, environmental information like the concentration of some nutrient or cytokine corresponds to the input to the channel, S, and the output can be quantified as some downstream molecular or phenotypic response, R [9]. The relevant quantity for measuring information transmission in signaling

networks is the mutual information:

$$I(S;R) \equiv \sum_{s\in S}\sum_{r\in R} p(s,r)\log\frac{p(s,r)}{p(s)\,p(r)}, \tag{5.1}$$

and it is ultimately a nonparametric measure of the correlation between the signal, S, and corresponding downstream response, R. The mutual information is generally calculated in units of bits, which comes from employing the base 2 logarithm in the calculation [26, 27]. Furthermore, estimation of the mutual information requires only that the signal variable and the response variable in question are measured; no underlying mechanistic knowledge of the signaling network is necessary. However, the mutual information depends on the signal distribution and thus does not necessarily reveal the underlying information transmission capabilities of the channel that are typically of primary interest. The relevant quantity for characterizing the limits of information transmission through some arbitrary signaling channel is the *channel capacity*, C, which is defined as the supremum of the mutual information over all possible probability distributions of the signal variable:

$$C \equiv \sup_{p_S(s)} I(S;R). \tag{5.2}$$

This quantity is a property of the channel itself, and it is the upper limit on the amount of information that can be transmitted through a channel [26, 27]. Note that our procedure only estimates the channel capacity, as numerical consideration of all signal distributions is computationally impossible (see Methods).

To date, there have been a number of studies examining the flow of information in intracellular signal transduction. The most notable was done by Andre Levchenko's group in 2011, in which they measured the information transmitted in the form of nuclear-localized NF-κB given some level of stimulation by TNF-α [9]. In this case, the response was measured at a particular point in time corresponding to the approximate peak in NF-κB localization at 30 minutes following stimulation. Despite the relative importance of this signaling network in governing cell-fate decisions, they ultimately found that the amount of information that can be transmitted by this network is

less than 1 bit, meaning that this particular signal-response pair is incapable of reliably making even binary decisions. Other studies have explored whether cells employ strategies to decrease the noise responsible for these seemingly low values (*e.g.* using dynamical trends as response to stimulus instead of a single point in time, or fold-change detection instead of concentration/-copy number of chemical species) [2, 28], or if noisy responses can be useful for responses at the level of cellular populations instead of individual cells (see Chapter 4). While these investigations have contributed greatly to our understanding of information transmission in specific signaling networks, a full understanding of the general properties of information transmission in signaling networks have not yet been realized. For instance, the majority of signaling networks whose ability to transmit information have been quantified exhibited channel capacities ranging between 0.5 and 2 bits of information on the level of the individual cell. It is currently unclear if these values are indicative of all signaling networks or if 2 bits of information represents an intrinsic upper limit on intracellular information transmission (see Chapter 4).

In this work, we characterized the limits of information transmission in intracellular signal transduction in order to develop a theoretical understanding of cellular decision-making in the context of information theory. In particular, we start by focusing on atomistic signaling motifs (*e.g.* binary ligand-receptor interactions) and then progress to slightly larger networks that still achieve a dynamic steady-state. In order to do this, we developed a framework for consistent comparison of information transmission in distinct systems. In the presence of only intrinsic biochemical noise, we show that smaller signaling motifs can readily achieve channel capacities of 5 bits of information transmission, and can exceed 6-7 bits in simple binary interaction, which is far more than has been observed experimentally. However, we do observe that in more complex motifs, such as kinase cascades, information content degrades as it transmitted through multiple stages of the network: cascades composed of 4 kinases transmit about 4 bits of information between signal and the most downstream kinase. Comparative estimates of the bounds of information transmission in these simple signaling motifs will then provide an intuitive basis for future work in characterizing cellular decision-making processes that occur via larger, more complex networks

(*e.g.* Wnt- or IGF-induced signaling). Finally, since this work examines specific signaling motifs, we also investigate whether or not certain biochemical trade-offs regulate or limit the flow of information through signaling networks in certain circumstances (*e.g.* saturation of enzymes in a covalent modification cycle tends to reduce information transmission). In addition to providing a platform for future work into quantifying information transmission in signaling networks, we expect that this work is fundamental to understanding why and how certain networks transmit specific levels of information.

5.2 Results

5.2.1 Framework

Prior methods used for data collection and estimation of information theoretic quantities did not consider specific factors that can impact the estimates [2, 9, 103, 132], and therefore cannot be deployed as is for consistent comparison between arbitrary signaling networks. As these factors are altered, such as the range or density of sampled signal values, the resulting mutual information or channel capacity estimates also change. In order to systematically investigate the upper limits of information transmission in cells, we developed a simple framework to control the factors that could impact estimation of the information theoretic quantities.

As a model system for this analysis, consider a signaling network defined only by a simple sigmoid function, commonly known as the Hill function in biochemistry:

$$R = R_{\min} + (R_{\max} - R_{\min}) \cdot \frac{S^n}{S^n + K^n} + \varepsilon, \tag{5.3}$$

where R and S denote response and signal, respectively, n controls the ultrasensitivity of the curve, and K is the signal value resulting in half-maximal response. This model also includes a noise term, ε, which is sampled from a Gaussian distribution with mean = 0, and some chosen standard deviation, σ. We chose this function for a number of reasons beyond its simplicity. This model

removes any assumptions about some underlying reaction network motif. Additionally, it turns out that many signaling motifs produce steady-state dose-response trends that are reminiscent of a sigmoid function, including phosphorylation cycles and kinase cascades [37, 87]. We proceeded to numerically sample signal and response pair from this model; the resulting dose-response data can be seen in Figure 5.1A for distinct levels of ultrasensitivity. Figure 5.1B shows the same data, but transformed to use indices for the signal in place of the raw values, where the minimum signal value for some arbitrary signal-response data set is assigned an index of 0 and the maximum is assigned an index of $N-1$ where N is the number of unique signal values sampled. This transformation conserves the underlying correlation between the signal and response (and thus does not alter mutual information calculations) and facilitates visual comparison between distinct data sets. Using data sets generated from this model we can investigate features of data collection that may impact the estimation of information theoretic quantities. Specifically, we examine how both the selection of distinct ranges of signal values and the number of sampled signal values impact channel capacity estimation.

Since collection of data spanning the entirety of signal space is obviously experimentally (and computationally) infeasible for any signaling network, we first characterized how the estimated channel capacity would change as the range of sampled signal values shifts. The majority of information should be within what we term the *increasing regime* or the *transition zone* of some arbitrary dose-response data set [114]. In most sigmoidal dose-response curves, the range of signal values corresponding to this transition zone spans values bounded by $S_{\text{min}} = S_{10}$ and $S_{\text{max}} = S_{90}$, which are the signal values resulting in 10% and 90% of the maximum response, respectively, after subtracting the baseline response [87, 133]. This eliminates a large range of signal space in which the response is not changing significantly with signal (Figure 5.1A). We examined whether the choice of these bounds impacts the estimation of information transmission and found that the effects are minimal (see Section D.1.2). To examine how shifting the window of signal space alters the channel capacity, we first fixed the number of uniformly sampled points to be 32. We define ΔS to be the shift in signal space from the mean signal value in the range of signal space

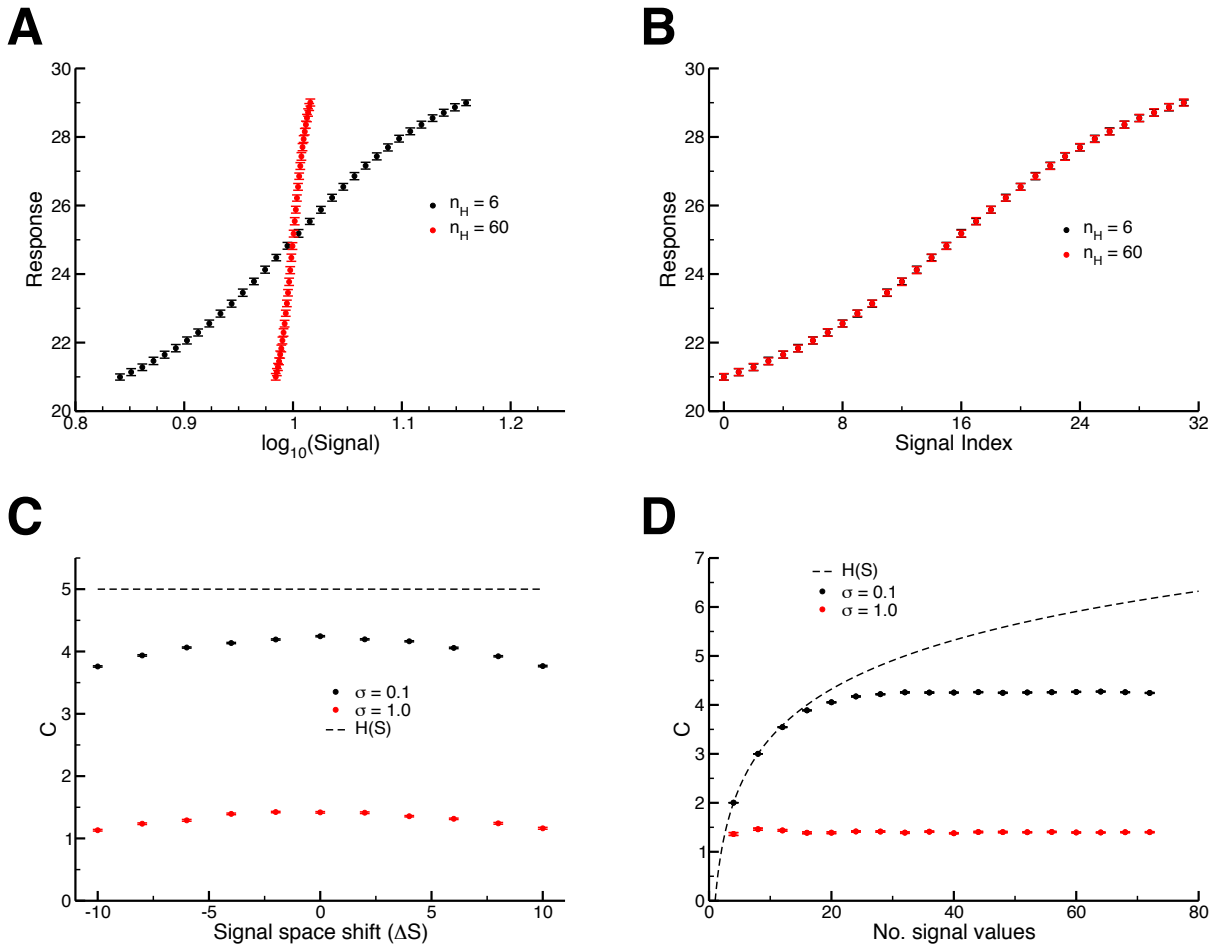


Figure 5.1: Characterizing the information in a simple signaling model. Error bars in this and subsequent figures denote 95% confidence intervals in the channel capacity estimation procedure. (A) Transition zone dose-response data from the Hill function model sampled with 32 signal values for two different values of n but identical levels of noise ($\sigma = 0.1$). (B) The data from panel (A) but with the signal values mapped to indices for comparative analysis. Note that while the transition zone shrinks with increasing ultrasensitivity, normalization to the width of the transition zone reveals the similarity of the two data sets (C) Channel capacity of the simple model for two levels of noise as a function of a shifting range of signal values. In both cases the maximum information transmission occurs when $\Delta S = 0$. (D) Channel capacity as a function of the density of signal values sampled in the transition zone. The minimum signal density for optimal information transmission depends on the noisiness of the channel. The entropy of the signal distribution is shown (black dotted line), denoting an upper bound to the channel capacity.

defined by the transition zone, where the unit is equal to the difference between signal values. Thus, when sampling signal values from the transition zone, $\Delta S = 0$. We observe, as expected, that shifting the range of sampled signal values away from the transition zone causes a reduction in information transmission for models exhibiting both low ($\sigma = 0.1$) and high ($\sigma = 1$) levels of noise (Figure 5.1C). This can be explained by the sigmoid shape of the dose-response curve: when the mean signal value is much less than or greater than K, mean responses from distinct signal values become more similar. Correlation, and thus information, between signal and response is lost.

Next, we characterized how the number of signal values uniformly sampled within the transition zone changes the information transmission. Generally, the number of signal values chosen to characterize a dose-response relationship is arbitrary, but this number can greatly impact estimation of the channel capacity. Since the mutual information (and by extension, the channel capacity) can be written as a difference of entropies:

$$I(S;R) = H(S) - H(S|R). \tag{5.4}$$

where $H(S)$ is the Shannon entropy of the signal, S, and $H(S|R)$ is the Shannon entropy of S conditioned on the response, R [27]. Thus, the entropy of the signal distribution, $H(S)$ is an upper bound on the mutual information. This limit can be reached with a uniform signal distribution when the signal values sampled produce pairwise disjoint response distributions, resulting in perfect information transmission between signal and response (Figure 5.1D). We can achieve higher mutual information by increasing the sampling density of signal space, however sampling nearby signal values can produce overlapping response distributions depending on how much noise is in the system, reducing information transmission efficiency. There is thus some sufficiently dense sampling of signal values in the transition zone beyond which the mutual information does not increase. In the case of our simple model, sampling 32 signal values is sufficient for reliable channel capacity estimation for our low noise model and sampling merely 8 is sufficient for the high noise model.

Using these systematic analyses, we can construct a methodology from which to reliably and consistently gain an understanding for how much information can be transmitted through some arbitrary signaling network. For data collection, the first step is to estimate the signal value resulting in half-maximal response and sample numerous signal values around this value. From this data set, a Hill function (or other appropriate function) can be parameterized through simple least-squares fitting techniques. This fit can then be employed as a guide to determine the signal values (S_{min} and S_{max}) that correspond to the bounds of the transition region for the network under investigation. From here, it is a simple matter to find the number of points in the transition region that "saturates" the information transmission as seen in Figure 5.1D. Throughout the remainder of this work, we will employed these strategies to systematically characterize the limits of information transmission in a number of common signaling motifs.

5.2.2 Information in binary interactions

We first focused on the simplest motif present in signal transduction networks: binary interaction. This physical association of two molecules can induce signal transmission through a variety of mechanisms, such as allostery or nucleation of higher-order macromolecular complexes. This type of interaction is ubiquitous; the quintessential example in the biology of signaling networks is the interaction between an extracellular ligand and a transmembrane receptor. Our model of the binary interaction is termed the *LT* model and we proceeded to examine the information transfer between the extracellular ligand concentration (the signal) and the ligand-bound form of the receptor (the response) at steady state. The model itself is composed of only two reactions: association of the ligand and receptor (L and T) to form a complex (B), and dissociation of the complex into its component parts. One benefit of this simple system is that we can analytically determine the bounds of the transition zone (in terms of total ligand concentration, L_T) by solving the standard binding isotherm for distinct levels of bound receptor:

$$S_{\text{min}} = L_{T,low} = \frac{0.1 \cdot B_{\text{max}} \left(0.1 \cdot B_{\text{max}} - K_D - T_T\right)}{0.1 \cdot B_{\text{max}} - T_T} \tag{5.5}$$

$$S_{\max} = L_{T,high} = \frac{0.9 \cdot B_{\max}\,(0.9 \cdot B_{\max} - K_D - T_T)}{0.9 \cdot B_{\max} - T_T}, \tag{5.6}$$

where $B_{\max} \equiv T_T$ is the maximum possible number of bound ligand-receptor complexes, K_D is the dissociation constant, and T_T is the total number of receptor molecules. The reactions in this model were defined using rule-based modeling languages, and we simulated this system using the associated exact stochastic simulators [14, 16] for a range of T_T values while keeping the quantity $\frac{T_T}{K_D}$ constant. Finally, since transformations of the data which preserve the underlying structure of the data do not alter the mutual information, we can map the signal values (which are discrete) to indices both for simpler labeling of the signal values and visualization of the data (Figure 5.2A).

Following the framework outlined previously, we varied the number of signal values sampled in the transition zone and estimated the channel capacity for each data set. As can be seen in Figure 5.2A, increases in receptor number correlate with less noise in response due to the stochastic effects of smaller copy numbers, resulting in increased information transmission (Figure 5.2B). Similar to our sigmoid model of signaling, we observe that there is some "saturating" density of signal values beyond which there is no increase in the information transmitted, and that this density depends on the variability in response of the system. At these saturating densities we observe strikingly high levels of information transmission (about 6 bits) compared to values previously calculated from experimental data sets [9].

In living systems, receptor and ligand numbers are kept constant through molecular turnover, which are themselves stochastic events, thus we introduced synthesis and degradation of the two molecule types, to examine the effect of this additional stochasticity on information transmission. This led to a slightly more complex, but still analytically tractable binding curve from which we determined the signal bounds of the transition zone (see Section D.2.1). The resulting dose-response data sets can be seen in Figure 5.2C, and the signal density-dependent channel capacity trends can be seen in Figure D.3. As anticipated, the additional variability introduced by molecular turnover reduced the information transmission. For saturated signal density, the loss in information was approximately 0.5 bits regardless of the receptor copy number. In general the trends observed for

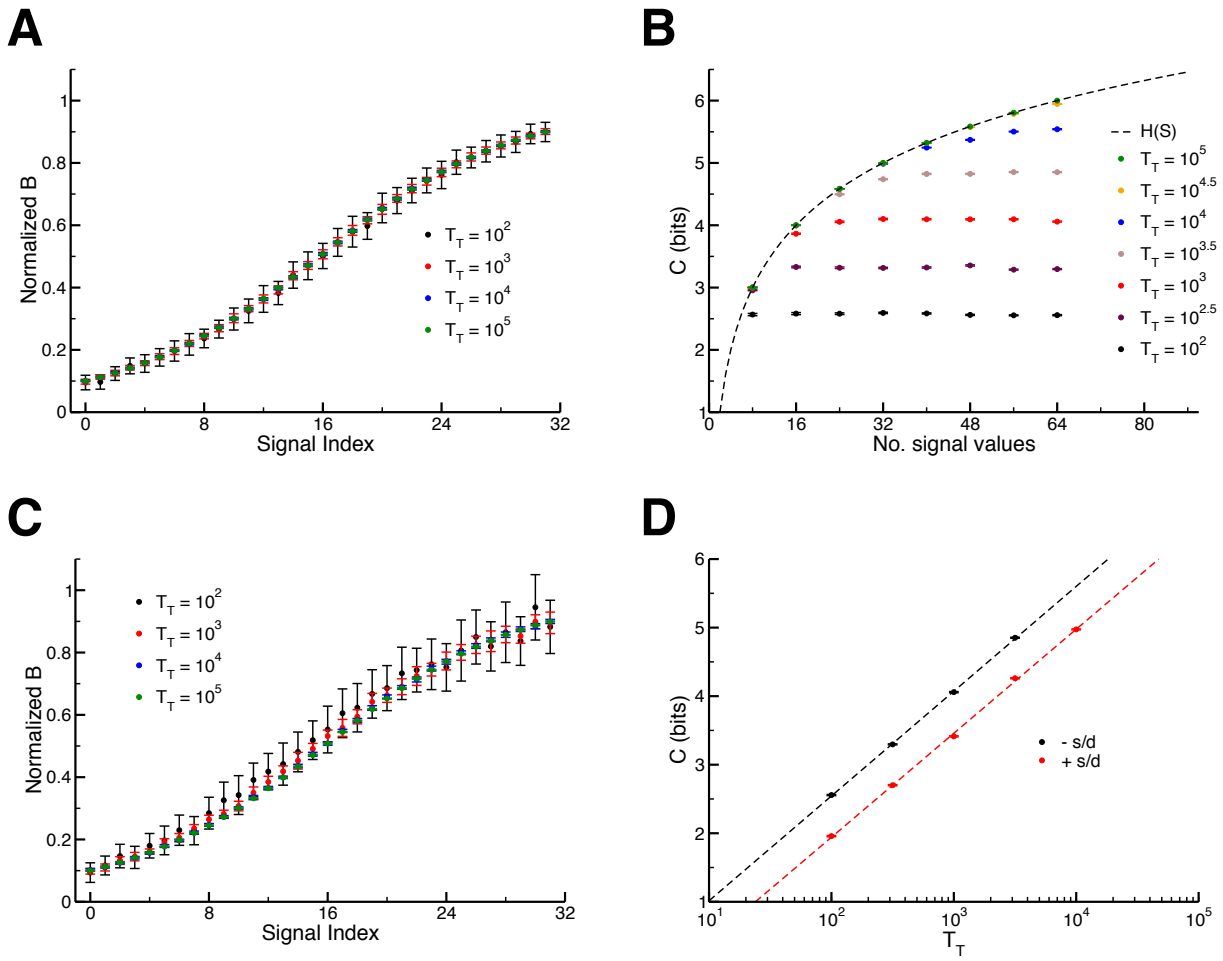


Figure 5.2: (A) The number of molecules in the systems (given by total receptor number, T_T) directly corresponds to the amount of noise in response (normalized bound receptor, B). (B) Similar to Figure 5.1, we observe "saturation" of information with sufficiently densely sampled signal values that depends on the level of noise in response. (C) The LT model that includes molecular turnover exhibits the same trends as in (A) but the relative level of noise is higher. (D) A log-linear relationship exists between the number of molecules in the system (inversely proportional to the variability in response) and the channel capacity. Note "$\pm$ s/d" in the panel denotes the presence or absence of synthesis and degradation (molecular turnover) in the model.

the model without synthesis and degradation were conserved in this model. As evident in Figure 5.2B, there exists a scaling relationship between the channel capacity of the system and the number of receptors. We found a significant log-linear relationship between the channel capacity (at saturating signal density) and receptor copy number in the range of receptor numbers we tested (Figure 5.2D). In other words, increasing receptor copy number to generate an additional 0.5 bits of information is offset by an order of magnitude cost in energy for receptor production. Energy cost aside, using this model to extrapolate to even larger numbers of receptors, we can estimate a hard upper bound on information transmission in cells. Since the ligand-receptor binding motif is required for virtually all signaling networks, its limitations are conferred on the rest of the network. We estimate that for an interaction with over 1 million receptors (an extreme upper bound, being nearly an order of magnitude larger than what has been experimentally realized for certain networks [134]), the amount of information that can be transmitted is approximately 8 bits (or 7.5 bits with molecular turnover). While this is unlikely to be realized *in vivo* since ligand-receptor interactions are generally only a part of larger signaling networks, this gives us a point of reference for understanding the limitations of information transmission through even the simplest signaling networks.

5.2.3 Information in futile cycles

We next focused on the standard chemical modification motif for signaling: Goldbeter & Koshland's covalent modification cycle (which we term the "GK loop") [87]. This model's kinetics have been thoroughly characterized mathematically for a number of operating regimes (*e.g.* saturated, unsaturated) and we can use this broad understanding to our advantage [86, 131, 135, 136]. While the model encompasses any type of enzymatic modification of substrate, our terminology will primarily reflect that of phosphorylation, where K, P, and W denote kinase, phosphatase and substrate, respectively. In our treatment of this model, the ratio of maximum velocities of the enzymes is the signal [86, 87, 135, 136]:

$$S \equiv \frac{k_{cat,K} \cdot [K_T]}{k_{cat,P} \cdot [P_T]}, \tag{5.7}$$

where $k_{cat,E}$ is the catalytic rate of the enzyme, E, and $[E_T]$ is the concentration of said enzyme. We vary this quantity by modifying $k_{cat,P}$, in order to more finely sample signal space; modifying the copy number of the kinase has undesired side effects on characterization of the transition zone (see Section D.3.1). The response, in this case is defined as the concentration of unbound, phosphorylated substrate [86, 87, 135]. A number of features of this motif are of interest for our information theoretic analysis; it is slightly more complex, having 6 total chemical species as opposed to 3 in the LT model, and it can exhibit ultrasensitivity when the enzymes are saturated. Our measure of saturation is $\frac{K_M}{W_T}$, where K_M is the Michaelis constant (equivalent for both enzymes) and W_T is total substrate concentration, and we varied this value from 10^{-2} to 10^2 (*i.e.* saturated to not saturated) logarithmically by modifying the binding rate of enzyme to substrate. Traditionally, altering the saturation of an enzyme is done by increasing the amount of total substrate, however we saw in the previous section that changes in copy number correspond to changes in information transmission. By changing K_M instead of W_T we control for this phenomenon while still being able to modify the saturation of the enzymes. We then applied our previously described fitting methodology in order to estimate the transition zone: we first sampled the response distributions for 20 points in signal space, taking care to capture the entire transition zone in this range of points, and then fit the data to a Hill function to determine $S_{\min}$ and $S_{\max}$. Finally, as with the LT model, we varied the copy numbers of the components in the system, while keeping the ratio of components fixed: $K_T : P_T : W_T = 1 : 1 : 100$.

Consistent with the LT model, altering the copy numbers of the signaling components served only to alter the variability of response to signal, and so a positive correlation again exists between copy number and information transmission. We also see, similar to the behavior in our initial, Hill function model, that increased ultrasensitivity (induced by enzyme saturation) tightens the signal bounds of the transition zone. If we again transform the discrete signal values to indices, we can visually compare how response distributions differ for a particular (relative) signal in the transition zone as seen in Figure 5.3A. Dose-response trends also emerge, revealing that saturated enzymes produce increased noise in response for signal values near the half-maximal signal value. This,

in turn reduces the amount of information present in the system, from nearly 5.5 bits in an unsaturated cycle to just above 3 bits (Figure 5.3B). Upon addition of synthesis and degradation to this model the information transmission is again generally reduced as the variability in response distributions is increased (Figure 5.3C), however the difference between unsaturated and saturated enzymes becomes exaggerated, varying between approximately 4 bits and less than 1 bit, respectively (Figure 5.3D). Thus, even a motif as simple as this GK loop can exist in a parameter regime in which information transmission is sufficiently low such that binary decisions are impossible to make reliably. It is therefore clear to see that enzyme kinetics, and the resulting ultrasensitivity of saturation in covalent modification cycles, have a large impact on information transmission, and must be tuned according to the particular cell-fate decisions they govern.

5.2.4 Information in kinase cascades

While useful for gaining an understanding of basic information transmission properties, these atomistic signaling motifs are rarely implemented in isolation. On the contrary, eukaryotic organisms and metazoans in particular exhibit increasingly complex cellular signaling networks in order to respond to environmental cues [90]. One of the most conserved network motifs is the kinase cascade, which is present in both simple eukaryotes, such as yeast [35], and complex multicellular life [137]. To examine how these complex networks transmit information via futile cycles and binary interactions, we constructed a set of rule-based models that embody the core of a kinase cascade. We have two similar, but distinct model types: one that employs a scaffold protein and one that does not [22], termed the *scaffold* and *solution* models, respectively. Both involve kinases which are sequentially phosphorylated and phosphatases assigned to dephosphorylate a specific kinase (to prevent saturation due to substrate-sharing) (Figure 5.4A & B) [86]. We also varied the number of successive kinase types in both the scaffold and solution models, and we refer to this number as the "depth", F, of the cascade. All parameters are based on those from the yeast MAPK signaling network and as such, the kinases are unsaturated [1]; additional details on model construction can be found in Section D.4.

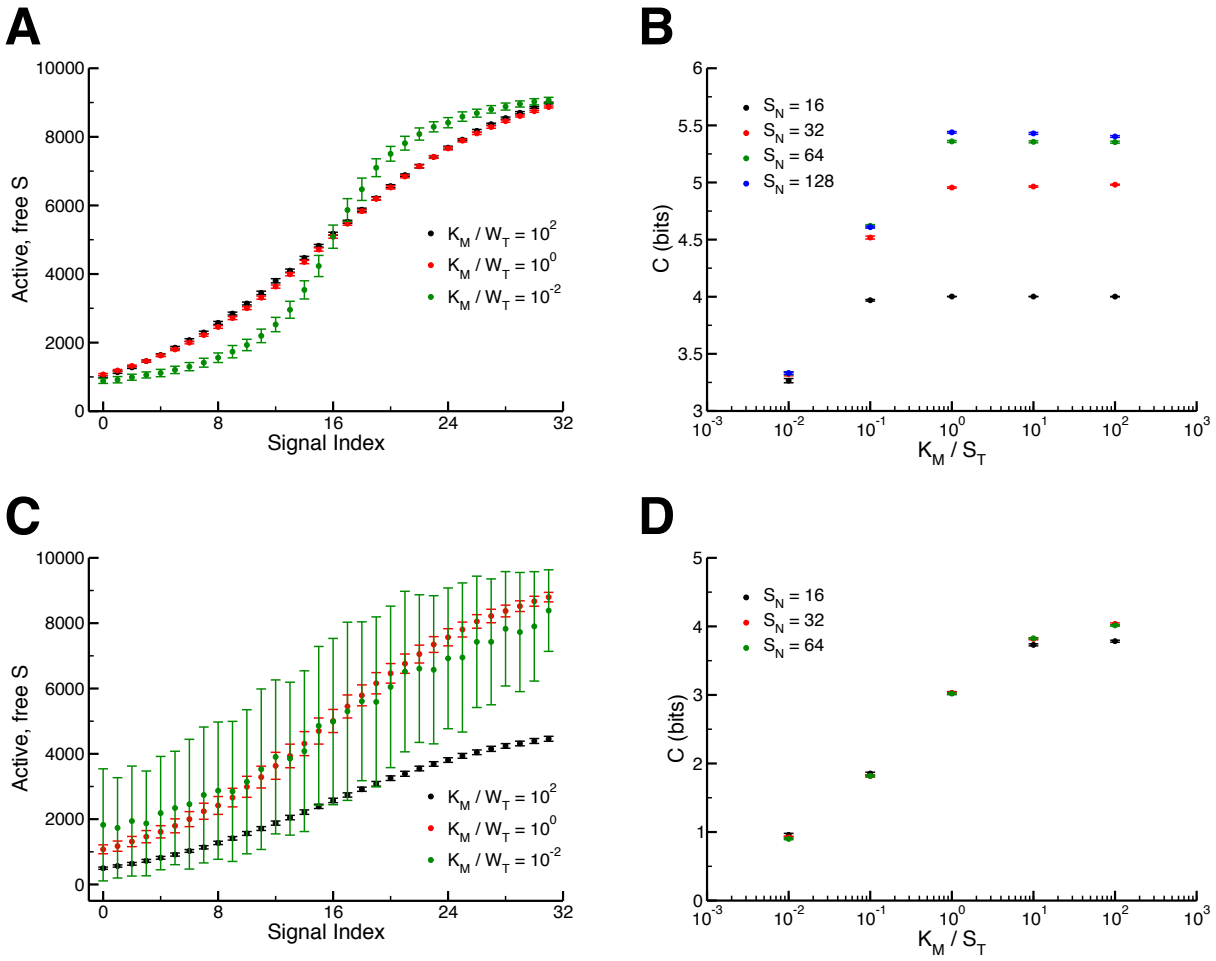


Figure 5.3: (A) Dose-response data sets for GK loops; S_N denotes number of signals sampled. Three different levels of enzyme saturation are shown; black is least saturated, green is most saturated. (B) Channel capacity as a function of enzyme saturation. As expected from (A), increased saturation results in lower information transmission due to high variability in the response distributions. (C) & (D) Identical to (A) and (B) but the model includes synthesis and degradation.

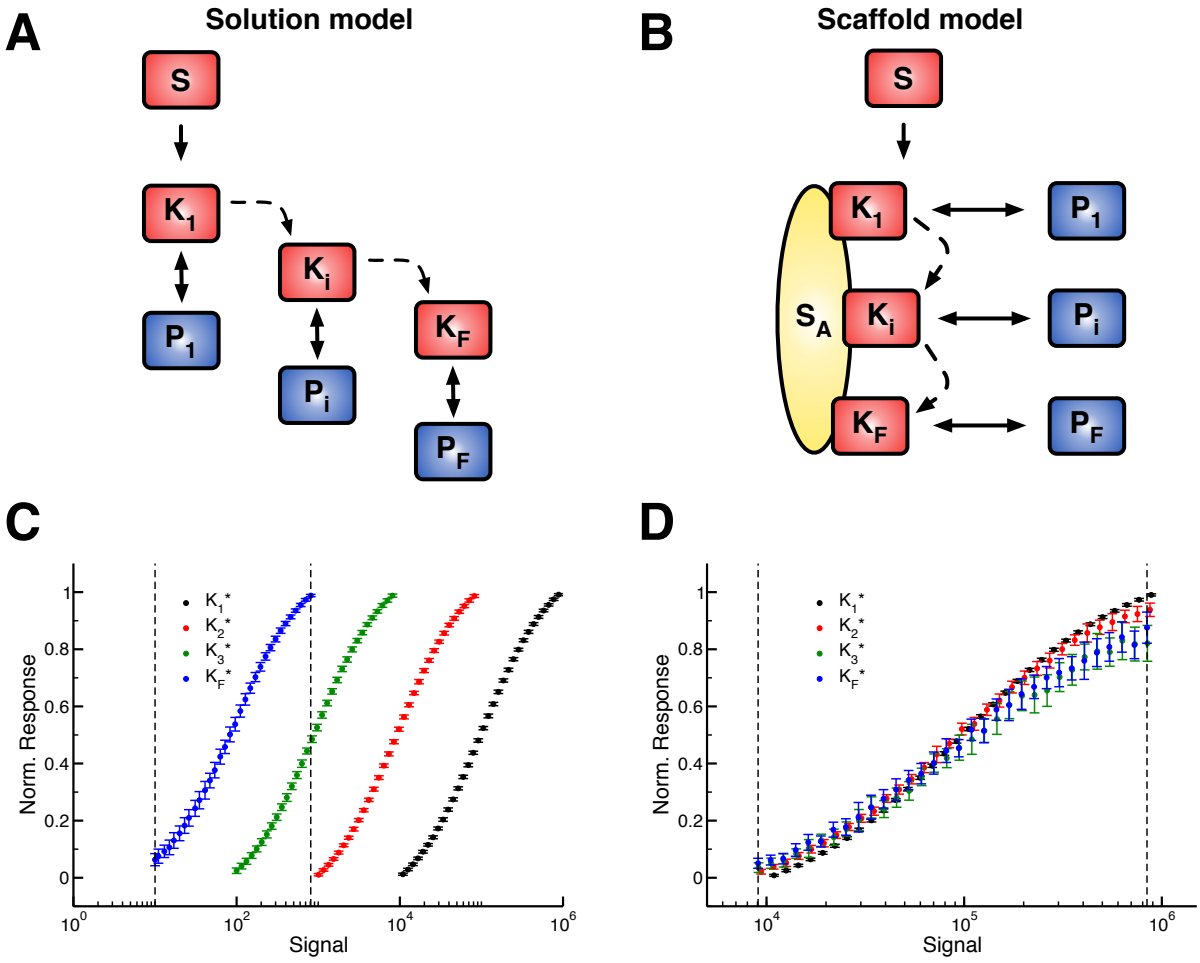


Figure 5.4: (A) Solution cascade model where activation of K_i only depends on binding active K_{i-1}. (B) Scaffold-based cascade model where activation of kinase K_i depends on having active K_{i-1} and K_i bound simultaneously to the scaffold, S. (C) Solution model dose-response trends generated using the VTZ approach (see text). The dotted lines denote the transition zone of the final kinase (which is used in the FTZ approach to generate dose-response data for all K_i). (D) As (C) but with the scaffold model.

We then generated dose-response data sets to examine the flow of information through these networks. In order to do this, we examined the information transmission between the signal and each kinase's activity in the cascade: $C(S,K_i)$; $0 \leq i \leq F$. With this data, we can begin to understand how both distinct interaction networks (*i.e.* the presence of a scaffold) and increases in the network size can influence information transmission. We calculated these values for cascades of depth 2, 3, and 4 in two different ways. In one approach, we calculated the transition zone bounds, $S_{\mathrm{min,i}}$ and $S_{\mathrm{max,i}}$, for each separate channel capacity estimation, $C(S,K_i)$, termed the variable transition zone (VTZ) approach. The other approach involves finding the transition zone bounds for the final kinase, $S_{\mathrm{max,F}}$ and $S_{\mathrm{min,F}}$, and using the resulting range of signal values for all estimations. This is the fixed transition zone (FTZ) approach. This latter approach is of significant interest since the final kinase in these types of cascades is typically responsible directly or indirectly for initializing some sort of transcriptional program that will govern behavioral changes in response to some stimulus [37].

From these dose-response data sets, we noticed a few key differences between the two models. First, the copy number of active, final kinase in the solution model for any given signal value is substantially higher than the scaffold model, reaching approximately 90% activation at the upper bound of the transition zone (Figure D.4). This is likely due to the increase in the number of conditions, and thus reaction events, required for activation to occur in the scaffold model; activation requires independent scaffold association for successive kinase and each kinases must bind the same scaffold, as opposed to a single binding interaction in the solution model (see Chapter 3). Second, the FTZ approach for the solution model samples a region of the upstream kinases' dose-response curves whose signal values are far smaller relative to the VTZ approach (*i.e.* $\Delta S < 0$), while the scaffold model exhibits minimal changes (Figure 5.4C & D). This preservation of dose-response shape through various stages of a signaling network has been previously identified as a feature of networks with scaffold proteins, and has been called *dose-response alignment* or DoRA [37].

Distinctions are also evident in the channel capacity estimations for each intermediate in the

cascade with $d = 4$ (Figure 5.5). With one exception, the information transmission at each stage in the scaffold model is lower than the corresponding stage in the solution model for the VTZ approach. This is likely due to the much lower magnitude of K_F response in the scaffold model (Figure D.4); stochastic effects in this portion of response space would be much greater than that in the solution model. We do observe monotonically decreasing channel capacity as information progresses through the cascade in both models, however there are contrasting trends: information transmission appears to drop more quickly but levels off at deeper stages in the scaffold model.

In the FTZ approach, a seemingly minor change in the protein interaction network induces relatively major behavioral differences between the dose-response relationships of these two network types. In the case of the scaffold model, the information between the signal and each intermediate is essentially equivalent to the VTZ case. In contrast, the channel capacity between the signal and early stages in the solution model is low. This is due to the signal space aligning poorly with the ideal transition zone of these early signaling species (*i.e.* the response to signal in early stages of the cascade is less sensitive than the later stages) (Figure 5.4). Most signal values sampled fall well below the presumed half-maximal signal value of the K_1 and K_2 intermediates' dose-response curve, and thus the channel capacity increases with the relative depth of the signaling species in question. At first glance, this may appear to violate the data processing inequality, which states that information content cannot be increased during transmission through a channel. However this is not the case in our model, since the signal and observed intermediates do not form a Markov chain [27]. While these results are undoubtedly subject to the parameter sets (*e.g.* the unsaturated enzymes allow much greater information transmission as observed in the GK loop model), the scaffold and solution model indicate that distinct configurations of signaling cascades' underlying interaction networks can impact information transmission. In general, these increasingly complex models of kinase cascades still exhibit a capacity for far more effective information transmission (>4 bits) than experimental data of cellular response to signal.

5.2.5 Information in realistic networks

With a basic understanding of how basic signaling motifs can influence information transmission in signaling networks, we can apply what we have learned to more realistic networks. Two-component signaling (TCS) systems present in bacteria are an ideal model system; they are stand-alone sensing networks that exhibit minimal crosstalk with other signaling systems [138, 139]. Additionally, the response in TCS is typically a transcription factor, and so this type of network embodies a complete signaling system, from extracellular stimuli, to alteration of gene expression levels. As the name suggests, this network involves two components: the Histidine Kinase and Response Regulator (HK & RR, respectively). This is one of the simplest and most ubiquitous in biology, and it has been extensively characterized experimentally and computationally, including some work on the fidelity of information transmission in the presence of crosstalk between HK-RR pairs [132, 138, 139]. Though similar to the futile cycle, this model differs significantly: the HK acts as both kinase and phosphatase to the RR, depending on its own phospho-state. Here, we adapted a previously implemented system of differential equations [139] to use rule-based stochastic simulations in order to introduce realistic levels of intrinsic noise into the system, and varied the existing kinetic parameters and protein copy numbers within ranges appropriate for bacterial systems [139].

As seen consistently throughout this work, scaling the protein copy numbers positively correlates with information transmission, and this network is no exception (Figure 5.6A & B). Furthermore, increased saturation again induces a decrease in information, due mainly to a reduction in the overall amount of active response regulator [139]. Most notably, however, we can use the parameters that most closely reflect existing experimental data and estimate the channel capacity of the HK-RR pair from which the original model was derived: Envz and OmpR, respectively [138, 140, 141] (Figure 5.6B, red box). Notably, these values are higher than nearly all experimentally characterized networks [9], showing that individual bacterial cells can, at least in principle, obtain relatively high quantities of information from extracellular stimuli.

We then turned our attention to a much more complex eukaryotic interaction network: re-

ceptor tyrosine kinase signaling cascades. The quintessential example is the epidermal growth factor (EGF) signaling network whose activity depends on ErbB2/HER transmembrane receptor dimerization. As mentioned before, a complete model of the system does not exist, however an experimentally validated rule-based model exploring the early events of this signaling cascade was employed to examine the network dynamics using exact stochastic simulation of all relevant species [43]. We adapted that model for use with our information theory framework to characterize how the information is transmitted through this network by examining the responses of various key signaling species. In particular, we focused on ligand-induced EGF receptor (EGFR) dimerization, autophosphorylated EGFR (which is required for recruitment of effector proteins), and active Sos (the downstream-most component in this model) (Figure 5.5C). This model contains nearly 200,000 EGFR molecules and we found that in order to accurately estimate the information transmission, we required a high signal density in the transition zone, sampling 256 distinct signal values from the transition zone in order to reach a "saturated" channel capacity estimation.

We observed high information transfer among the initial steps of the cascade: EGFR dimerization in response to EGF stimulation produced nearly 6.5 bits of information, close to the upper bound estimated from the LT model. From this point, the stochasticity of the interactions and lower copy number of other components, such as Sos, reduces the information transmission (Figure 5.6D). However, information transmission through the entirety of the network was greater than 3 bits. This is quite high compared to experimentally determined values, however it is important to note that Sos recruitment is by no means the final step of the cascade *in vivo*. Sos is then responsible for activating the MAPK pathway in metazoan signaling, and we have seen that information transfer can vary significantly, depending on the kinetics of the kinase cascade (Figure 4.3B & D). It is enough, though, to see that even with a model containing moderate signaling complexity (over 350 signaling species as opposed to 3 in the LT model or 6 in the GK model), reliable information transmission is possible.

5.3 Discussion

This work begins to address a fundamental question in the study of signal transduction and cellular-decision making: why do signaling networks transmit specific amounts of information? Here, we focused on characterizing the limits of information transmission through signaling networks with the goal of providing context for information theoretic values estimated from experimental data. We found that models of simple signaling motifs, as well as larger, more realistic networks, are capable of transmitting substantially higher amounts of information than has been estimated experimentally; to date, the highest information transmission estimated for individual signaling networks in eukaryotic cells (that we are aware of) is less than 2.5 bits. There are a few possibilities as to why this might be the case. First, it is possible that the networks that have been examined were simply those that did not require high information transmission, and others exist that transmit much more. We predict that networks with the highest information transmission will be those whose corresponding cell fate decision either exists in continuous space or is a categorical variable with high entropy, and which requires precise decision-making on the part of individual cells to maintain organism viability (*e.g.* three-dimensional spatial resolution in chemotaxis or differentiation of pluripotent stem cells). Second, the observed low levels of information in *in vivo* signal transduction could be a limitation of extrinsic noise [4–6]. Our models only include the intrinsic randomness of biochemical reaction events, however other environmental factors can contribute to overall variability in response to signal, including noise in the signal distribution itself. Finally, we showed in prior work that low information transmission (even values below 1 bit of information) can be useful when transmitting information to cellular populations is paramount. We found that there exists a fundamental trade-off between information transmission to single cells and cellular populations, and that there is some optimal level of noise to maximize population-level information transmission given certain conditions (Chapter 4). Regardless of the reason for low observed information transmission, we were able to examine how common signaling events, like chemical modification, could be potential mechanisms for the regulation of information transmission.

In order to compare information transmission values from data obtained by simulation of mul-

tiple signaling motifs, we introduced a novel framework for the consistent application of information theoretic concepts to systems biology. While quantities such as the mutual information do not depend on the underlying structure of the reaction network between signal and response, they do depend on how joint signal-response space is characterized (Figure 5.1), meaning that dose-response data from distinct networks must be obtained using a standard methodology. The framework developed here can be applied both to simulated and experimental data sets and is crucial to make relevant comparisons of estimated values, such as those from different cell types or lines, organisms, or even networks with recombinant proteins (which could prove to be useful in the construction of *de novo* networks via synthetic biology). The results shown in this work are a strong example of the power of such a tool; we were able to characterize the upper limits of information transmission in networks of varying size and interconnectivity through stochastic simulation of rule-based models. The channel capacities for the initial atomistic signaling motif models that we examined (binary, physical interaction between macromolecules and chemical modification of macromolecules via enzymes) are generally integrated into more complex network architecture and can therefore provide perspective for analysis of larger networks, which are common in metazoan cells. For example, since examination of the covalent modification cycle revealed that enzyme saturation reduces information transmission, we restricted our analysis of the larger cascade models (which employ a variant of this motif) to those with unsaturated enzymes since we were primarily concerned with the upper limits of information transmission. In general, this framework provides a basis for understanding and comparing how various features (*e.g.* molecular copy numbers, kinetic parameters) influence the quantity of information transmitted through signaling networks, and it can serve as a standard for future application of information theory to quantification of information in signal transduction.

However, we expect that this framework is just a starting point for broader application of information theory to systems biology. To fully understand how information flows through a network, the dynamics of the network must be considered. While a useful starting point, quantities like the mutual information represent levels of information at particular points in time. In this work,

we focused on networks whose components exhibited distinct steady-state responses, however the response of some signaling networks, such as those that exhibit perfect adaptation , are not well-characterized by steady-state response [142]. It is clear that the framework presented here must be further developed, since cellular decision-making rarely waits until steady state is reached (at least on the molecular level) [37]. Previous examples of information theory applied to *in vivo* signaling networks circumvented this problem by defining a particular point in time to measure response, a logical solution when there exists some particular time at which an interesting event occurs, such as peak nuclear NF-κB localization [9]. Another possibility is to define the response as a set of multiple time points from a dynamical trend, however the *in vivo* mechanistics of such a response are not immediately clear [2]. More complex quantities, such as the transfer entropy [128, 130], have been derived to examine how information flows through a system over time eliminating the need for making arbitrary choices about when to take measurements. Adapting these quantities for use with high resolution time course data will undoubtedly elucidate additional principles of information flow in signal transduction. Extracellular signals and their corresponding cellular responses lie at the core of what cells have evolved to do: adapt to changing environmental conditions by altering their phenotype. Ultimately, we expect that development and application of systematic approaches such as this will form the basis for what becomes a rigorous theory of information transmission through signaling networks.

5.4 Methods

5.4.1 Mutual information estimation

Our calculation is loosely based on those previously employed by [9] as defined in Chapter 4. In essence, we apply resampling strategies to eliminate the inherent upward bias of estimating the mutual information from finite data sets [116, 119, 143]. We first define a number of bins in signal and response space in order to partition the coupled signal and response values. With this binning, we construct a contingency table (*i.e.* a two-dimensional histogram) and estimate the

mutual information. We perform this procedure for resamplings of the data using smaller sample sizes in order to extrapolate the mutual information to infinite sample size using simple linear regression. We then iterate over both the number of bins used for the calculation and a predefined set of unimodal and bimodal weights to modify the signal distribution to maximize the mutual information and gain an estimate of the channel capacity. We further control for artificial inflation of the estimate due to high numbers of bins by calculating the mutual information for a randomized data set given some number of bins and checking to see if the information is not significantly different from 0. In order to estimate the channel capacity, we weight the signal distribution of the data using a set of unimodal and bimodal probability distributions.

5.4.2 Model simulation

We employed the Doob-Gillespie stochastic simulation algorithm to generate all dynamical data in this work [89]. Both the Kappa [15, 16] and BioNetGen [14] rule-based modeling languages and associated software packages (which have been shown to be equivalent in their simulation of stochastic models [1]) were used for model construction.

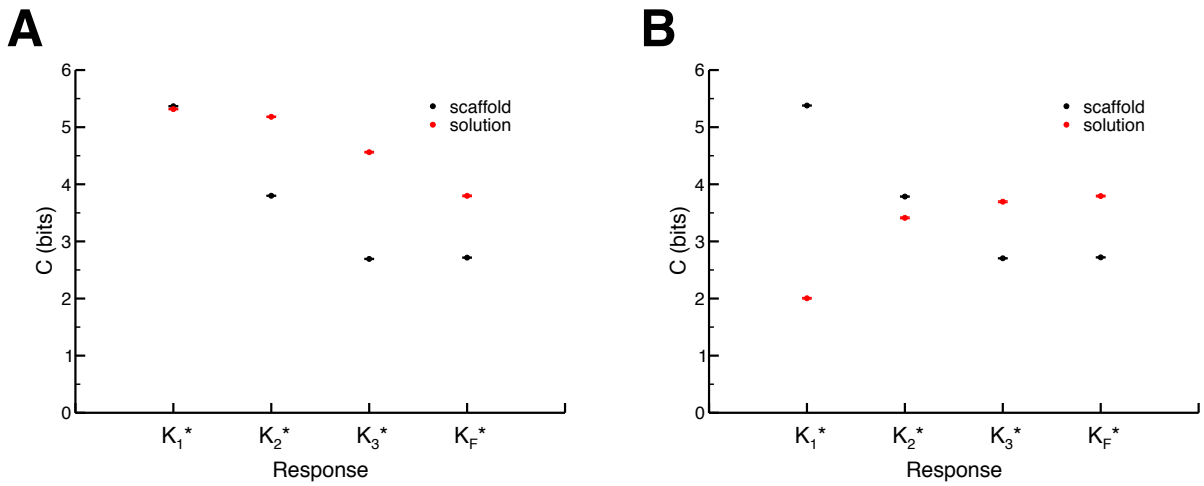
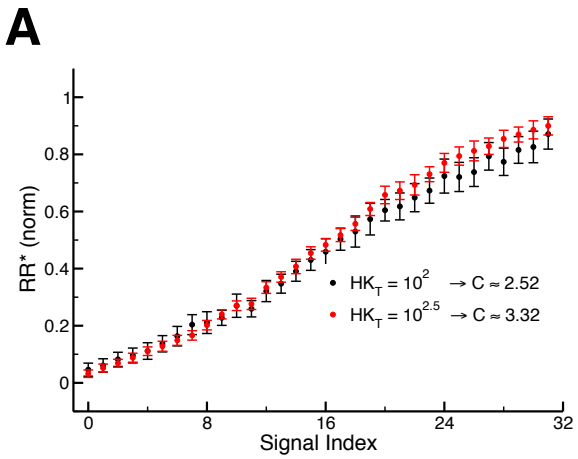


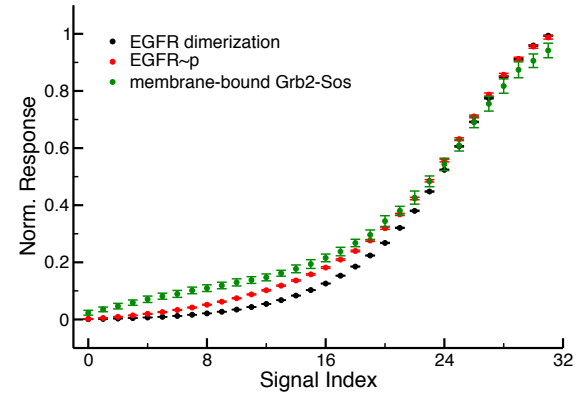
Figure 5.5: (A) & (B) Channel capacity as a function of cascade depth for data VTZ (A) and FTZ (B) models. Distinctive behavior arises with relatively minor differences between the models; the trends in the scaffold model are qualitatively identical for both VTZ and FTZ models, whereas the solution model exhibits strikingly different behavior due to the increased sensitivity to signal of its downstream components.



B

HK copy number

K_M / W_T	5×10^1	$5 \times 10^{1.5}$	5×10^2
10^1	3.36	4.13	4.78
10^0	2.52	3.32	3.98
10^{-1}	1.54	2.26	2.99
10^{-2}	1.20	1.92	2.67



D

Signal	Response	C (bits)
EGF	EGFR dimers	6.44 ± 0.003
EGF	phos. EGFR	5.80 ± 0.004
EGF	active Sos	3.58 ± 0.004

Figure 5.6: (A) Dose-response data for distinctly parameterized TCS motifs. (B) Channel capacity values for various parameterizations of the TCS model. As with the GK model, the range of values varies with both saturation and component copy number. The numbers boxed in red represent the models whose parameters are closest to experimental data from the EnvZ/OmpR TCS that is responsible for osmoregulation in bacteria such as *E. coli*. (C) Dose-response data for key observables in a rule-based model of the early events in EGFR signaling. (D) Channel capacities for the observables shown in (C). Note that while the data set in (C) is composed of 32 signal values, the channel capacities were calculated using 256 signal values to reach sufficiently dense signal value sampling in the transition zone. The necessity of high signal density is due to the large number of EGFR molecules in the system, a phenomenon we observed in the LT model (Figure 5.2)

Chapter 6

Conclusion

Heterogeneity in metazoan cells has begun to emerge as prominent topic of consideration for experimental and theoretical systems biologists over the last decade, both in terms of compositional heterogeneity in protein complex assembly in signaling networks [1, 10–12, 43, 44], and the general variability, or noise, present in cells [4, 5, 9, 28, 144]. Here, we used novel approaches to characterize the effects of this heterogeneity on signal transduction in order to develop an understanding of how cellular decision-making can proceed in the face of uncertainty. With mathematical tools developed specifically for addressing these questions, we were able to elucidate general properties of signaling that should inform future experimental and theoretical investigations into cellular decision-making.

In Chapter 2, we showed that a model of the yeast pheromone signaling network with minimal assumptions exhibited extreme combinatorial complexity, and thus compositional heterogeneity, but was still able to reliably reproduce experimentally observed trends [1]. This shows that signaling through *pleiomorphic ensembles* is a feasible mode of signal transduction [11]. Furthermore the presence of combinatorial inhibition in this network [20, 21], but not in a distinct *machine*-like model of signaling (Figure 2.6), provides indirect evidence of the existence of ensemble signaling. However, in depth and technically challenging experimental work will likely need to be employed for direct detection of ensemble signaling. Specifically, the transient nature of protein complexa-

tion implied in ensemble signaling makes *in vivo* characterization difficult, however the existence of super-resolution microscopy could allow detection of such complexes in real time. Ultimately, this work showed that the means by which cells assemble signaling complexes must be realized for accurate predictive modeling of signal transduction.

These results also revealed that certain features (*i.e.* combinatorial inhibition) may be limited to a particular mode of assembly. We further showed in Chapter 3 that a number of dynamical or dose-response features are distinct between machine- or ensemble-like signaling, or between scaffold-based signaling paradigms and those without scaffolds. Of note is the presence of *increased* signal amplification, defined as the ratio of activity between the final kinase and initial kinase for a cascade of arbitrary depth, in the machine-like signaling paradigm. Prior reviews and speculation on the behavior of scaffold proteins posited that they (absent any consideration of the conditions required for assembly) would always *reduce* signal amplification [22, 23]. We observed a slight reduction in amplification for the ensemble model in comparison to the scaffoldless solution model, however the amplification was still present, showing that formal investigation of signaling dynamics can reveal non-intuitive features related to combinatorial complexity in protein interaction networks. These investigations into structural and compositional heterogeneity among complexes and its effects on signaling dynamics promoted the idea put forth by Gerhart & Kirschner that ensemble- and machine-like complexation could serve alternative evolutionary roles [25]. Specifically, the independence of signaling components in ensembles could facilitate robustness in the rewiring of a signaling network, potentially preventing catastrophic failure upon mutation of individual components. [13, 76, 78]. Machines, on the other hand, already exist in the cell, and have proven their usefulness (in evolutionary terms) by maintaining core processes, such as translation and protein degradation, that are conserved across all forms of life.

These signaling networks and core processes, however, are all subject to various forms of biochemical noise, both intrinsic and extrinsic [5]. Noise is generally considered undesirable, as it degrades the information that can be sent between a signal and response; for individual cells, sufficiently high levels of noise could produce a seemingly arbitrary boundary between distinct cell-fate

decisions. However Chapter 4 describes how noise can be beneficial in certain circumstances, such as the need to control proportions of cells in a population that undertake some decision in response to a given signaling factor. In particular, we found that there is an optimal level of noise that maximizes information flow to cellular populations in a simple model of signaling (Figure 4.3). These apparently high levels of noise are unlikely to be a biophysical limit of signaling networks. Other systems that evolved to control cellular behaviors critical for individuals cells (*e.g.* chemotaxis, yeast mating) exhibit notably higher information transmission than other previously characterized networks [9]. We posit that noise is therefore a regulated quantity, and that cells could theoretically implement mechanisms to tune noise in certain signaling networks, depending on how it best optimizes organismal fitness. This has important implications for the evolution of multicellular life, since essential features of multicellularity (*i.e.* tissue homeostasis) requires management of population-level behaviors.

The exact mechanisms for this are unknown, but in Chapter 5 we have begun to develop a theoretical basis for understanding the limits of information transmission in various signaling motifs. For information transmission through networks at steady-state, we were able to estimate an upper bound of 7-8 bits in the presence of only intrinsic noise (Figure 5.2). This provides a context for understanding existing channel capacity calculations from experimental data. Our characterization of commonly found motifs in larger networks provides an intuition for how information flows through more complex networks. This suggests several mechanisms for how cells could reduce the flow of information (as in Chapter 4): saturation of enzymes in covalent modification cycles are capable of generating sufficient noise to lower information transmission to about 1 bit (Figure 5.3). Evolution of a network with multiple saturated cycles could easily allow increased information flow to cellular populations by generating noisy responses at the level of the individual cell.

Cellular heterogeneity, traditionally considered an obstacle for cells to overcome, can have non-intuitive and potentially desirable effects for signal transduction. As shown in this work, investigation of cellular decision-making in the presence of compositional heterogeneity and biochemical noise can be performed through formal, systematic analyses involving mathematical modeling and

numerical simulation without undue simplification. In fact, models whose purpose is to examine signaling dynamics must not write off such heterogeneity, or they risk incorrect prediction of cellular behavior. We expect that the approaches developed in this work will spur the adoption of the necessary modeling techniques for further elucidation of general properties of information transmission in cells in the presence of cellular heterogeneity. Ultimately, the characterization of these properties will undoubtedly advance both our ability to manipulate these signaling systems for various biomedical applications and, more conceptually, our understanding of the role of signal transduction in the context of cellular and multicellular evolution.

References

[1] Suderman R, Deeds EJ (2013) Machines vs. Ensembles: Effective MAPK Signaling through Heterogeneous Sets of Protein Complexes. PLoS Computational Biology 9: e1003278.

[2] Selimkhanov J, Taylor B, Yao J, Pilko A, Albeck J, et al. (2014) Accurate information transmission through dynamic biochemical signaling networks. Science (New York, NY) 346: 1370–1373.

[3] Alberts B (1998) The Cell as a Collection of Protein Machines. Cell 92: 291–294.

[4] Elowitz MB (2002) Stochastic Gene Expression in a Single Cell. Science (New York, NY) 297: 1183–1186.

[5] Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. PNAS 99: 12795–12800.

[6] Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK (2009) Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature 459: 428–432.

[7] Feinerman O, Veiga J, Dorfman JR, Germain RN, Altan-Bonnet G (2008) Variability and Robustness in T Cell Activation from Regulated Heterogeneity in Protein Levels. Science (New York, NY) 321: 1081–1084.

[8] Paulsson J, Ehrenberg M (2000) Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. Physical Review Letters 84: 5447–5450.

[9] Cheong R, Rhee A, Wang CJ, Nemenman I, Levchenko A (2011) Information transduction capacity of noisy biochemical signaling networks. Science (New York, NY) 334: 354–358.

[10] Hlavacek WS, Faeder JR, Blinov ML, Perelson AS, Goldstein B (2003) The complexity of complexes in signal transduction. Biotechnology and Bioengineering 84: 783–794.

[11] Mayer BJ, Blinov ML, Loew LM (2009) Molecular machines or pleiomorphic ensembles: signaling complexes revisited. Journal of Biology 8: 81.

[12] Deeds EJ, Krivine J, Feret J, Danos V, Fontana W (2012) Combinatorial complexity and compositional drift in protein interaction networks. PloS One 7: e32032.

[13] Sato PM, Yoganathan K, Jung JH, Peisajovich SG (2014) PLOS Biology: The Robustness of a Signaling Complex to Domain Rearrangements Facilitates Network Evolution. PLoS Biology 12: e1002012.

[14] Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M, et al. (2006) Rules for Modeling Signal-Transduction Systems. Science's STKE 2006: re6.

[15] Danos V, Feret J, Fontana W, Harmer R, Krivine J (2007) Rule-Based Modelling of Cellular Signalling. In: Caires L, Vasconcelos VT, editors, CONCUR-2007 — Concurrency Theory, Springer Berlin Heidelberg, volume 4703 of *Lecture Notes in Computer Science*. pp. 17-41.

[16] Danos V, Feret J, Fontana W, Krivine J (2007) Scalable Simulation of Cellular Signaling Networks. In: Shao Z, editor, Programming Languages and Systems, Springer Berlin Heidelberg, volume 4807 of *Lecture Notes in Computer Science*. pp. 136-157.

[17] Gillespie DT (2007) Stochastic Simulation of Chemical Kinetics. Annual Review of Physical Chemistry 58: 35–55.

[18] Sneddon MW, Faeder JR, Emonet T (2010) Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. Nature Publishing Group 8: 177–183.

[19] Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature Cell Biology 440: 631–636.

[20] Levchenko A, Bruck J, Sternberg PW (2000) Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. PNAS 97: 5818-23.

[21] Chapman SA, I AR (2009) Quantitative effect of scaffold abundance on signal propagation. Molecular Systems Biology 5: 313.

[22] Good MC, Zalatan JG, Lim WA (2011) Scaffold proteins: hubs for controlling the flow of cellular information. Science (New York, NY) 332: 680–686.

[23] Burack WR, Shaw AS (2000) Signal transduction: hanging on a scaffold. Current Opinion in Cell Biology 12: 211–216.

[24] Patterson JC, Klimenko ES, Thorner J (2010) Single-cell analysis reveals that insulation maintains signaling specificity between two yeast MAPK pathways with common components. Science Signaling 3: ra75.

[25] Gerhart J, Kirschner M (2007) The theory of facilitated variation. PNAS 104: 8582–8589.

[26] Shannon CE (1948) A Mathematical Theory of Communication. Bell System Technical Journal 27: 379–423.

[27] Cover TM, Thomas JA (1991) Elements of Information Theory. New York, NY, USA: Wiley-Interscience.

[28] Lee REC, Walker SR, Savery K, Frank DA, Gaudet S (2014) Fold Change of Nuclear NF-κB Determines TNF-Induced Transcription in Single Cells. Molecular Cell 53: 1-13.

[29] Marques AJ, Palanimurugan R, Matias AC, Ramos PC, Dohmen RJ (2009) Catalytic Mechanism and Assembly of the Proteasome. Chemical Reviews 109: 1509–1536.

[30] Bashan A, Yonath A (2008) Correlating ribosome function with high-resolution structures. Trends in Microbiology 16: 326–335.

[31] Korostelev A, Noller HF (2007) The ribosome in focus: new structures bring new insights. Trends in Biochemical Sciences 32: 434–441.

[32] Kiel C, Serrano L (2011) Challenges ahead in signal transduction: MAPK as an example. Current Opinion in Biotechnology 23: 1–10.

[33] Qi S, Pang Y, Hu Q, Liu Q, Li H, et al. (2010) Crystal Structure of the Caenorhabditis elegans Apoptosome Reveals an Octameric Assembly of CED-4. Cell 141: 446–457.

[34] McClean MN, Mody A, Broach JR, Ramanathan S (2007) Cross-talk and decision making in MAP kinase pathways. Nature Genetics 39: 409–414.

[35] Chen RE, Thorner J (2007) Function and regulation in MAPK signaling pathways: lessons learned from the yeast Saccharomyces cerevisiae. Biochimica et Biophysica Acta 1773: 1311–1340.

[36] Shao D, Zheng W, Qiu W, Ouyang Q, Tang C (2006) Dynamic studies of scaffold-dependent mating pathway in yeast. Biophysical Journal 91: 3986–4001.

[37] Yu RC, Pesce CG, Colman-Lerner A, Lok L, Pincus D, et al. (2008) Negative feedback that improves information transmission in yeast signalling. Nature 456: 755–761.

[38] Andersson J, Simpson DM, Qi M, Wang Y, Elion EA (2004) Differential input by Ste5 scaffold and Msg5 phosphatase route a MAPK cascade to multiple outcomes. The EMBO Journal 23: 2564–2576.

[39] Danos V, Feret J, Fontana W, Krivine J (2008) Abstract interpretation of cellular signalling networks. In: Verification, Model Checking, and Abstract Interpretation, Springer Berlin Heidelberg, volume 4905 of *Lecture Notes in Computer Science*. pp. 83–97.

[40] Thomson M, Gunawardena J (2009) Unlimited multistability in multisite phosphorylation systems. Nature 460: 274–277.

[41] Onuchic JN, Wolynes PG (2004) Theory of protein folding. Current Opinion in Structural Biology 14: 70–75.

[42] Shakhnovich E (2006) Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chemical Reviews 106: 1559–1588.

[43] Blinov ML, Faeder JR, Goldstein B, Hlavacek WS (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. Bio Systems 83: 136–151.

[44] Faeder JR, Blinov ML, Goldstein B, Hlavacek WS (2005) Combinatorial complexity and dynamical restriction of network flows in signal transduction. Systems Biology 2: 5–15.

[45] Mody A, Weiner J, Ramanathan S (2009) Modularity of MAP kinases allows deformation of their signalling pathways. Nature Cell Biology 11: 484–491.

[46] Babu M, Vlasblom J, Pu S, Guo X, Graham C, et al. (2012) Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. Nature 489: 585–589.

[47] Villalobos V, Naik S, Bruinsma M, Dothager RS, Pan MH, et al. (2010) Dual-color click beetle luciferase heteroprotein fragment complementation assays. Chemistry & Biology 17: 1018–1029.

[48] Ridgeway WK, Millar DP, Williamson JR (2012) Quantitation of ten 30S ribosomal assemblyintermediates using fluorescence triplecorrelation spectroscopy. PNAS 109: 13614–13619.

[49] Hoskins AA, Friedman LJ, Gallagher SS, Crawford DJ, Anderson EG, et al. (2011) Ordered and dynamic assembly of single spliceosomes. Science (New York, NY) 331: 1289–1295.

[50] Huang B, Babcock H, Zhuang X (2010) Breaking the diffraction barrier: super-resolution imaging of cells. Cell 143: 1047–1058.

[51] Betzig E, Patterson GH, Sougrat R, Lindwasser OW, Olenych S, et al. (2006) Imaging intracellular fluorescent proteins at nanometer resolution. Science (New York, NY) 313: 1642–1645.

[52] Manley S, Gillette JM, Patterson GH, Shroff H, Hess HF, et al. (2008) High-density mapping of single-molecule trajectories with photoactivated localization microscopy. Nature Publishing Group 5: 155–157.

[53] Rust MJ, Bates M, Zhuang X (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nature Methods 3: 793–795.

[54] Yi TM, Kitano H, Simon MI (2003) A quantitative characterization of the yeast heterotrimeric G protein cycle. PNAS 100: 10764–10769.

[55] Bhattacharyya RP, Reményi A, Good MC, Bashor CJ, Falick AM, et al. (2006) The Ste5 scaffold allosterically modulates signaling output of the yeast mating pathway. Science (New York, NY) 311: 822–826.

[56] Yablonski D, Marbach I, Levitzki A (1996) Dimerization of Ste5, a mitogen-activated protein kinase cascade scaffold protein, is required for signal transduction. PNAS 93: 13864–13869.

[57] Thomson TM, Benjamin KR, Bush A, Love T, Pincus D, et al. (2011) Scaffold number in yeast signaling system sets tradeoff between system output and dynamic range. PNAS 108: 20265–20270.

[58] Danos V, Feret J, Fontana W, Harmer R, Krivine J (2008) Rule-Based Modelling, Symmetries, Refinements. In: Fisher J, editor, Formal Methods in Systems Biology, Springer Berlin Heidelberg, volume 5054 of *Lecture Notes in Computer Science*. pp. 103-122.

[59] Ghaemmaghami S, Huh W, Bower K (2003) Global analysis of protein expression in yeast. Nature 425: 737-741.

[60] Madura K, Varshavsky A (1994) Degradation of G alpha by the N-end rule pathway. Science (New York, NY) 265: 1454–1458.

[61] Klinke DJ (2009) An empirical Bayesian approach for model-based inference of cellular signaling networks. BMC Bioinformatics 10: 371.

[62] Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, et al. (2009) Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. Molecular Systems Biology 5: 239.

[63] Gonzalez OR, Küper C, Jung K, Naval PC, Mendoza E (2007) Parameter estimation using Simulated Annealing for S-system models of biochemical networks. Bioinformatics (Oxford, England) 23: 480–486.

[64] Creamer MS, Stites EC, Aziz M, Cahill JA, Tan CW, et al. (2012) Specification, annotation, visualization and simulation of a large rule-based model for ERBB receptor signaling. BMC Systems Biology 6: 107.

[65] Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, et al. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127: 635–648.

[66] Williamson JR (2008) Cooperativity in macromolecular assembly. Nature Chemical Biology 4: 458–465.

[67] Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ (2007) Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. Proteomics 7: 944–960.

[68] van Dongen SM (2000) Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht.

[69] Bray D, Lay S (1997) Computer-based analysis of the binding steps in protein complex formation. PNAS 94: 13493–13498.

[70] Dembo M, Goldstein B (1978) Theory of equilibrium binding of symmetric bivalent haptens to cell surface antibody: application to histamine release from basophils. Journal of Immunology 121: 345–353.

[71] Saiz L, Vilar JM (2006) Stochastic dynamics of macromolecular-assembly networks. Molecular Systems Biology 2: 24.

[72] Oberdorf R, Kortemme T (2009) Complex topology rather than complex membership is a determinant of protein dosage sensitivity. Molecular Systems Biology 5: 253.

[73] Deeds EJ, Bachman JA, Fontana W (2012) Optimizing ring assembly reveals the strength of weak interactions. PNAS 109: 2348-2353.

[74] Uversky VN (2002) Natively unfolded proteins: A point where biology waits for physics. Protein Science 11: 739–756.

[75] Levy Y, Wolynes PG, Onuchic JN (2004) Protein topology determines binding mechanism. PNAS 101: 511–516.

[76] Peisajovich SG, Garbarino JE, Wei P, Lim WA (2010) Rapid Diversification of Cell Signaling Phenotypes by Modular Domain Recombination. Science (New York, NY) 328: 368–372.

[77] Kitano H (2004) Biological robustness. Nature Reviews Genetics 5: 826–837.

[78] Bashor CJ, Helman NC, Yan S, Lim WA (2008) Using Engineered Scaffold Interactions to Reshape MAP Kinase Pathway Signaling Dynamics. Science (New York, NY) 319: 1539–1543.

[79] Good M, Tang G, Singleton J, Reményi A, Lim WA (2009) The Ste5 scaffold directs mating signaling by catalytically unlocking the Fus3 MAP kinase for activation. Cell 136: 1085–1097.

[80] Locasale JW, Shaw AS, Chakraborty AK (2007) Scaffold proteins confer diverse regulatory properties to protein kinase cascades. PNAS 104: 13307–13312.

[81] Park SH, Zarrinpar A, Lim WA (2003) Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. Science (New York, NY) 299: 1061–1064.

[82] Saito H (2010) Regulation of cross-talk in yeast MAPK signaling pathways. Current Opinion in Microbiology 13: 677–683.

[83] Zarrinpar A, Bhattacharyya RP, Nittler MP, Lim WA (2004) Sho1 and Pbs2 act as coscaffolds linking components in the yeast high osmolarity MAP kinase pathway. Molecular Cell 14: 825–832.

[84] Yang J, Hlavacek WS (2011) Scaffold-mediated nucleation of protein signaling complexes: Elementary principles. Mathematical Biosciences 232: 164–173.

[85] Wang Y, Dohlman HG (2004) Pheromone signaling mechanisms in yeast: a prototypical sex machine. Science (New York, NY) 306: 1508–1509.

[86] Rowland MA, Fontana W, Deeds EJ (2012) Crosstalk and competition in signaling networks. Biophysical Journal 103: 2389–2398.

[87] Goldbeter A, Koshland DE (1981) An amplified sensitivity arising from covalent modification in biological systems. PNAS 78: 6840–6844.

[88] McCaffrey G, Clay F, Kelsay K (1987) Identification and regulation of a gene required for cell fusion during mating of the yeast Saccharomyces cerevisiae. Molecular and Cellular Biology 7: 2680-2690.

[89] Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry 81: 2340-2361.

[90] Kirouac DC, Saez-Rodriguez J, Swantek J, Burke JM, Lauffenburger DA, et al. (2012) Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. BMC Systems Biology 6: 29.

[91] Bardwell L (2006) Mechanisms of MAPK signalling specificity. Biochemical Society Transactions 34: 837.

[92] Danos V, Feret J, Fontana W, Harmer R, Hayman J, et al. (2012) Graphs, Rewriting and Pathway Reconstruction for Rule-Based Models. In: D'Souza D, Kavitha T, Radhakrishnan J, editors, IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2012). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, volume 18 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 276–288.

[93] Bardwell L (2005) A walk-through of the yeast mating pheromone response pathway. Peptides 26: 339–350.

[94] Chen JY, Lin JR, Cimprich KA, Meyer T (2012) A Two-Dimensional ERK-AKT Signaling Code for an NGF-Triggered Cell-Fate Decision. Molecular Cell 45: 196–209.

[95] Balázsi G, van Oudenaarden A, Collins JJ (2011) Cellular Decision Making and Biological Noise:From Microbes to Mammals. Cell 144: 910–925.

[96] Sebolt-Leopold JS, Herrera R (2004) Targeting the mitogen-activated protein kinase cascade to treat cancer. Nature Reviews Cancer 4: 937–947.

[97] Fallahi-Sichani M, Honarnejad S, Heiser LM, Gray JW, Sorger PK (2013) Metrics other than potency reveal systematic variation in responses to cancer drugs. Nature Chemical Biology 9: 708–714.

[98] Flusberg DA, Roux J, Spencer SL, Sorger PK (2013) Cells surviving fractional killing by TRAIL exhibit transient but sustainable resistance and inflammatory phenotypes. Molecular Biology of the Cell 24: 2186–2200.

[99] Albeck JG, Burke JM, Aldridge BB, Zhang M, Lauffenburger DA, et al. (2008) Quantitative Analysis of Pathways Controlling Extrinsic Apoptosis in Single Cells. Molecular Cell 30: 11–25.

[100] Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nature Reviews Genetics 6: 451–464.

[101] Ahrends R, Ota A, Kovary KM, Kudo T, Park BO, et al. (2014) Controlling low rates of cell differentiation through noise and ultrahigh feedback. Science (New York, NY) 344: 1384–1389.

[102] Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, et al. (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature 510: 363-369.

[103] Mehta P, Goyal S, Long T, Bassler BL, Wingreen NS (2009) Information processing and signal integration in bacterial quorum sensing. Molecular Systems Biology 5: 325.

[104] Hartwell LH, Culotti J, Pringle JR, Reid BJ (1974) Genetic control of the cell division cycle in yeast. Science (New York, NY) 183: 46–51.

[105] Doncic A, Falleur-Fettig M, Skotheim JM (2011) Distinct interactions select and maintain a specific cell fate. Molecular Cell 43: 528–539.

[106] Doncic A, Skotheim JM (2013) Feedforward regulation ensures stability and rapid reversibility of a cellular state. Molecular Cell 50: 856–868.

[107] Doncic A, Atay O, Valk E, Grande A, Bush A, et al. (2015) Compartmentalization of a bistable switch enables memory to cross a feedback-driven transition. Cell 160: 1182–1195.

[108] Levchenko A, Nemenman I (2014) Cellular noise and information transmission. Current Opinion in Biotechnology 28: 156–164.

[109] Cohen-Saidon C, Cohen AA, Sigal A, Liron Y, Alon U (2009) Dynamics and variability of ERK2 response to EGF in individual living cells. Molecular Cell 36: 885–893.

[110] Bao XR, Fraser IDC, Wall EA, Quake SR, Simon MI (2010) Variability in G-protein-coupled signaling studied with microfluidic devices. Biophysical Journal 99: 2414–2422.

[111] Coppey M, Boettiger AN, Berezhkovskii AM, Shvartsman SY (2008) Nuclear trapping shapes the terminal gradient in the Drosophila embryo. Current Biology 18: 915–919.

[112] Janetopoulos C, Firtel RA (2008) Directional sensing during chemotaxis. FEBS Letters 582: 2075–2085.

[113] R Core Team (2014) R: A Language and Environment for Statistical Computing. The R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

[114] Laughlin S (1981) A simple coding procedure enhances a neuron's information capacity. Zeitschrift für Naturforschung Section C: Biosciences 36: 910–912.

[115] Brenner N, Bialek W, de Ruyter van Steveninck R (2000) Adaptive rescaling maximizes information transmission. Neuron 26: 695–702.

[116] Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: detecting and evaluating dependencies between variables. Bioinformatics (Oxford, England) 18 Suppl 2: S231–40.

[117] Paninski L (2003) Estimation of Entropy and Mutual Information. Neural Computation 15: 1191–1253.

[118] Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Physical Review E 69: 066138.

[119] Strong S, Koberle R, de Ruyter van Steveninck R, Bialek W (1998) Entropy and Information in Neural Spike Trains. Physical Review Letters 80: 197–200.

[120] Miller M, Hafner M, Sontag E, Davidsohn N, Subramanian S, et al. (2012) Modular Design of Artificial Tissue Homeostasis: Robust Control through Synthetic Cellular Heterogeneity. PLoS Computational Biology 8: e1002579.

[121] Sacan A, Ferhatosmanoglu H, Coskun H (2008) CellTrack: an open-source software for cell tracking and motility analysis. Bioinformatics (Oxford, England) 24: 1647–1649.

[122] Burov S, Tabei SMA, Huynh T, Murrell MP, Philipson LH, et al. (2013) Distribution of directional change as a signature of complex dynamics. PNAS 110: 19689–19694.

[123] Shahrezaei V, Swain PS (2008) The stochastic nature of biochemical networks. Current Opinion in Biotechnology 19: 369–374.

[124] Friedman N, Cai L, Xie XS (2006) Linking stochastic dynamics to population distribution: an analytical framework of gene expression. Physical Review Letters 97: 168302.

[125] Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. Nature 440: 358–362.

[126] Kepler TB, Elston TC (2001) Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophysical Journal 81: 3116–3136.

[127] Debnath J, Muthuswamy SK, Brugge JS (2003) Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods (San Diego, Calif) 30: 256–268.

[128] Wollstadt P, Martínez-Zarzuela M, Vicente R, Díaz-Pernas FJ, Wibral M (2014) Efficient transfer entropy analysis of non-stationary neural time series. PloS One 9: e102833.

[129] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1: S7.

[130] Schreiber T (2000) Measuring information transfer. Physical Review Letters 85: 461–464.

[131] Levine J, Kueh HY, Mirny L (2007) Intrinsic Fluctuations, Robustness, and Tunability in Signaling Cycles. Biophysical Journal 92: 4473–4481.

[132] Lyons SM, Prasad A (2012) Cross-talk and information transfer in mammalian and bacterial signaling. PloS One 7: e34488.

[133] Ventura AC, Bush A, Vasen G, Goldín MA, Burkinshaw B, et al. (2014) Utilization of extracellular information before ligand-receptor binding reaches equilibrium expands and shifts the input dynamic range. PNAS 111: E3860–E3869.

[134] Lauffenburger DA, Linderman JJ (1996) Receptors. Models for Binding, Trafficking, and Signalling. International Journal of Biochemistry and Cell Biology 12: 1418.

[135] Rowland MA, Harrison B, Deeds EJ (2015) Phosphatase specificity and pathway insulation in signaling networks. Biophysical Journal 108: 986–996.

[136] Gomez-Uribe C, Verghese GC, Mirny LA (2007) Operating Regimes of Signaling Cycles: Statics, Dynamics, and Noise Filtering. PLoS Computational Biology 3: e246.

[137] Kiel C, Serrano L (2009) Cell type-specific importance of ras-c-raf complex association rate constants for MAPK signaling. Science Signaling 2: ra38.

[138] Batchelor E, Goulian M (2003) Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system. PNAS 100: 691–696.

[139] Rowland MA, Deeds EJ (2014) Crosstalk and the evolution of specificity in two-component signaling. PNAS 111: 5550–5555.

[140] Cai SJ, Inouye M (2002) EnvZ-OmpR interaction and osmoregulation in Escherichia coli. Journal of Biological Chemistry 277: 24155–24161.

[141] Qin L, Yoshida T, Inouye M (2001) The critical role of DNA in the equilibrium between OmpR and phosphorylated OmpR mediated by EnvZ in Escherichia coli. PNAS 98: 908–913.

[142] Muzzey D, Gómez-Uribe CA, Mettetal JT, van Oudenaarden A (2009) A systems-level analysis of perfect adaptation in yeast osmoregulation. Cell 138: 160–171.

[143] Treves A, Panzeri S (1995) The Upward Bias in Measures of Information Derived from Limited Data Samples. Neural Computation 7: 399–407.

[144] Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. Nature 467: 167–173.

[145] Danos V, Laneve C (2004) Formal molecular biology. Theoretical Computer Science 325: 69–110.

[146] Feret J, Danos V, Krivine J, Harmer R, Fontana W (2009) Complex Systems: From Chemistry to Systems Biology Special Feature: Internal coarse-graining of molecular systems. PNAS 106: 6453–6458.

[147] Powell CD (2003) Chitin scar breaks in aged Saccharomyces cerevisiae. Microbiology 149: 3129–3137.

[148] Inouye C, Dhillon N, Durfee T, Zambryski PC, Thorner J (1997) Mutational analysis of STE5 in the yeast Saccharomyces cerevisiae: application of a differential interaction trap assay for examining protein-protein interactions. Genetics 147: 479–492.

[149] Maleri S, Ge Q, Hackett EA, Wang Y, Dohlman HG, et al. (2004) Persistent activation by constitutive Ste7 promotes Kss1-mediated invasive growth but fails to support Fus3-dependent mating in yeast. Molecular and Cellular Biology 24: 9221–9238.

[150] Mahanty SK, Wang Y, Farley FW, Elion EA (1999) Nuclear shuttling of yeast scaffold Ste5 is required for its recruitment to the plasma membrane and activation of the mating MAPK cascade. Cell 98: 501–512.

[151] Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, et al. (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science (New York, NY) 287: 873–880.

[152] Wang Y, Christley S, Mjolsness E, Xie X (2010) Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent. BMC Systems Biology 4: 99.

[153] Leeuw T, Wu C, Schrag JD, Whiteway M, Thomas DY, et al. (1998) Interaction of a G-protein beta-subunit with a conserved sequence in Ste20/PAK family protein kinases. Nature 391: 191–195.

[154] Xu S, Kamath MV, Capson DW (1993) Selection of partitions from a hierarchy. Pattern Recognition Letters 14: 7 - 15.

[155] Albeck JG, Burke JM, Spencer SL, Lauffenburger DA, Sorger PK (2008) Modeling a Snap-Action, Variable-Delay Switch Controlling Extrinsic Cell Death. PLoS Biology 6: e299.

Appendix A

Appendix for Chapter 1

A.1 Yeast pheromone model

The model, rules and parameters used in our simulations were developed based on a number of sources, most notably the annotated online model found at `http://www.yeastpheromonemodel.org` that is written in the BioNetGen rule-based modeling language (BNGL) [18]. Numerous additional rules (including those regarding the nuclear shuttling of Ste5) were derived from equations and mechanisms present in Shao *et al.*'s ODE model [36]. Our final model, written in the Kappa language [15, 39, 58, 145, 146], has a total of 232 rules and all but one follow mass-action kinetics. The rules themselves are provided in an additional supplementary Kappa file.

A.1.1 Initial conditions

The initial conditions for our model (protein copy numbers) were derived from [57, 59] and the online model (OM) with the exception of the phosphatases for Ste11 and Ste7, which are unknown and estimated. All eight gene agents have a copy number of 1. Also it is important to note that the number of pheromone (*i.e.* α-factor) agents varied among simulations. Our dose-response curves clearly required different levels of pheromone stimulation, however for the drift calculations we used a concentration of 100 nM (10000 molecules).

Protein	Copy number
Pheromone	varies
Ste2	10000
Gpa1/Ste4 complex	10000
Gpa1 monomer	5000
Sst2	2500
Ste20	4200
Ste5	1680
Ste11	3500
Ste7	960
Fus3	20400
Kss1	20800
Msg5	38
Ptp	1270
Ste11-phosphatase (Mekkp)	1750
Ste7-phosphatase (Mekp)	1750
Dig1	3409
Dig2	1184
Ste12	1390

Table A.1: Protein copy numbers

A.1.2 Rate parameters

In the following sections we will discuss the numerous rate parameters in our model of the yeast pheromone signaling system, and so, for clarity's sake, we have classified the parameters into three categories:

1. directly observed in yeast (*D*)
2. indirectly inferred from similar systems (*e.g.* ERK phosphorylation) or previously used in other models (*I*)
3. unknown and estimated (*U*)

In total, 17 (7%) of the rate parameters in the ensemble model were directly observed, 158 (68%) were inferred, and 57 (25%) were unknown and estimated.

In the main text, we mention that 111 parameters were identified as potentially influencing dynamical trends seen in experimental data. We varied these parameters and ultimately determined

that 25 of them had a strong impact on the observed trends; these numbers are shown in red in the following subsections. Of these 25, 1 was directly observed, 22 were inferred, and 2 were unknown. We identified these parameters through trial and error, hypothesizing which parameters govern certain experimentally characterized trends (if no such hypothesis currently exists) and modifying them to better match said trends. For example, the Ste4 synthesis rate likely controls the long-term increase after the initial peak in the G protein activation time-course plot, as seen in the main text, Fig. 2.2B [54]. We therefore altered this rate over numerous iterations until our simulations accurately reproduced the observed experimental trend. The subsequent table lists those parameters that were modified to fit experimental observation in addition to the specific trends they affect. It is important to note that, due to a lack of model identifiability, there may be other parameters that alter the experimental trends in question; we chose these parameters simply due to their large relative influence on the dynamics of observables.

In order to maintain a biologically realistic model, our modifications to these rates were confined to reasonable limits. We allowed variation of approximately one order of magnitude for parameters inferred from related systems and completely unknown parameters were estimated according to the following table:

In subsections 1.3 - 1.8 there are tables of the associated rate parameters used in the model and their sources (OM for online model). The leftmost column contains the interaction or reaction. These are condensed descriptions of the actual rules, which are explicitly defined in the Kappa rule file. As such, there may be numerous rates for a particular interaction or reaction indicating that the rate differs in specific contexts (*e.g.* binding partners, phosphorylation states). The center-left column contains the rate parameter(s) and the center-right column mentions the model or other source from which it was derived. The parameter's category (D, I or U) is seen in the rightmost column. Note that as these are stochastic rates; the parameters for bimolecular reactions thus depend on the volume used for the yeast cell [12, 57], which may not be identical between models.

Rate Parameter	Trend	Cat.
Sst2/Ste2 assoc.	controls slope of decline in G protein activation curve following the initial peak	I
Sst2/Ste2 dissoc.	same as above	I
GTP hydrolysis	controls the time of the initial G protein activation peak	I
Gpa1 degradation	controls G protein levels and the relative levels of G protein activation	D
Ste4 degradation	same as above	I
Gpa1/Ste4 dimer deg.	same as above	I
Ste4/Ste20 dissoc.	determines the time course of Ste4-Ste20 binding	I
Ste4/Ste5 dissoc. (2 rates)	controls membrane localization of Ste5	I
Fus3 phos. (4 rates)	controls time and absolute value of Fus3 activation	I
Ste11 degradation	controls Ste11 levels and thus active Fus3 levels by extension (also involved with negative feedback and thus dose-response (DR) alignment)	I
Fus3/Msg5 dissoc. (4 rates)	controls peak Fus3 activation and DR trends	I
Fus3/Ptp dissoc. (2 rates)	same as above	I
Fus3 dephos. by Ptp	controls peak Fus3 activation and DR trends	I
Ste12/Gpa1_gene assoc.	controls rate of Gpa1 synthesis (G protein activation) and thus the relative level of active G protein compared to the initial peak	U
Gpa1 synthesis	same as above	I
Ste12/Ste4_gene assoc.	same as above	U
Ste4 synthesis	same as above	I

Table A.2: Influential rate parameters

Reaction Type	Parameter Range
K_D for cytosolic protein-protein interactions	10^2 nM
K_D for nuclear-localized protein-protein interactions	$10^1 - 10^3$ nM
k_{cat} for catalysis reactions	$10^{-1} - 10^1$ s^{-1}
k_{deg} for degradation reactions	$10^{-4} - 10^{-2}$ s^{-1}
k_{synth} for synthesis reactions	$10^{-1} - 10^1$ s^{-1}

Table A.3: Parameter variation range

A.1.3 G-protein cycle

The initial events of the pheromone response network involve the extracellular pheromone binding to the G-protein coupled receptor, Ste2. Since we are implementing a stochastic model and our rates require a specific volume, we define our extracellular volume to be V_{ext} = 166 fL and our intracellular volume to be V_{int} = 19.3 fL [147]. Briefly, the G-protein cycle passes the extracellular signal (the presence of pheromone) to the MAPK cascade via a nucleotide exchange mechanism. Our understanding of this process in yeast comes from [36, 54]. Upon activation of the G-protein coupled receptor (Ste2), the α subunit of the heterotrimeric G-protein (Gpa1), bound to Ste2, exchanges its bound GDP for GTP, thereby inducing dissociation from the $\beta\gamma$ complex (Ste4-Ste18, subsequently referred to just as Ste4). This allows Ste4, which is tethered to the membrane via Ste18 (implicit in our model) to recruit Ste5 and induce the MAPK cascade (see Sections A.1.5 & A.1.8). The GTPase-Activating Protein (GAP), Sst2, is able to bind Ste2, and the resulting colocalization of Sst2 and Gpa1 via Ste2 catalyzes GTP hydrolysis. Sst2 thus acts as a negative regulator, and enables Gpa1 to rebind Ste4 [54].

Note that numerous association rates among binary interactions in the G-protein cycle (Ste2/Gpa1 binding) are significantly higher than those later in the cascade (*e.g.* Ste5/Ste7 binding). This is due to the membrane association and localization of certain proteins (*e.g.* Ste4) and results in a higher apparent association rate (derived in the online model).

A.1.4 Ensemble MAPK cascade

Upon dissociation from Gpa1, Ste4 can engage in a number of different interactions. It can of course rebind Gpa1, but it can also bind the p21-activated kinase (PAK) Ste20 and recruit the scaffold protein, Ste5, to initiate the MAPK cascade [35]. Ste5 in turn must bind Ste11, a MAPKKK, simultaneously with a Ste4 bound to a Ste20 and form a 4-member complex in order for the Ste20 to phosphorylate Ste11. Though this complex is required for signal transduction, no experimental evidence suggests any particular binding order for creation of this tetramer. Upon Ste5 dimerization, Ste11 can then cross-phosphorylate the MAPKK, Ste7, on the opposite Ste5 [148]. Active

Interaction or Reaction	Rate parameter(s)	Source	Cat.
Pheromone/Ste2 interaction	3×10^{-5} molec^{-1}s^{-1}	OM, [54]	D
	0.015 s^{-1}	OM, [54]	D
Ste2/Gpa1 assoc.	0.001725 molec^{-1}s^{-1}	derived	U
Ste2/Gpa dissoc.	0.15 s^{-1}	OM	I
	0.03 s^{-1}	N/A	U
Gpa1/Ste4 assoc.	0.001725 molec^{-1}s^{-1}	derived	I
	8.595×10^{-8} molec^{-1}s^{-1}	derived	I
Gpa1/Ste4 dissoc.	7.5 s^{-1}	N/A	U
	1.5 s^{-1}	N/A	U
GDP $\rightarrow$ GTP	0.15 s^{-1}	[54]	D
Sst2/Ste2 interaction	8.595×10^{-4} molec^{-1}s^{-1}	derived	I
	0.15 s^{-1}	derived	I
GTP $\rightarrow$ GDP	0.015 s^{-1}	OM	I
	0.015 s^{-1}	OM	I
	1.5 s^{-1}	OM	I
	1.5 s^{-1}	OM	I
Sst2/MAPK assoc. (2 MAPKs)	8.595×10^{-5} molec^{-1}s^{-1}	derived	I (2)
Sst2/MAPK dissoc.	1.5 s^{-1}	OM	I (2)
	0.75 s^{-1}	OM	I (2)
	0.75 s^{-1}	OM	I (2)
	0.375 s^{-1}	OM	I (2)
Sst2 phosphorylation	1.5 s^{-1}	OM	I
Sst2 dephos.	0.00087 s^{-1}	[36]	I
Ste2 endocytosis	0.0004 s^{-1}	[54]	D
	0.003 s^{-1}	[54]	D
G-protein degradation	4.95×10^{-5} s^{-1}	OM	D
	4.95×10^{-5} s^{-1}	OM	I
	3.3×10^{-5} s^{-1}	OM	I
Sst2 degradation	0.0004 s^{-1}	OM	D
	0.0006 s^{-1}	OM	D

Table A.4: G-protein cycle interactions

Ste7 bound to Ste5 can then phosphorylate the MAPK, Fus3 [35]. As mentioned in the main text, only those dependencies explicitly demonstrated experimentally are implemented in our ensemble model (*e.g.* Ste5 need not necessarily be bound to Ste4 for Ste7 to phosphorylate Fus3). Note that interactions/reactions mentioning “MAPK" mean that the particular event could involve either Fus3 *or* Kss1 (a Fus3 paralog).

A.1.5 MAPK cascade regulation

Our model includes two MAPK phosphatase agents, Msg5 and Ptp (the latter of which represents two *in vivo* phosphatases, Ptp2 and Ptp3) that dephosphorylate Fus3 [36]. We also included individual phosphatases for Ste11 and Ste7, though the proteins which play this role *in vivo* remain to be experimentally characterized [36]. Also, this model employs a number of feedback mechanisms [36]. Active Fus3 can phosphorylate Ste11 (on a domain distinct from its activation domain) and Sst2, tagging them for degradation, which is modeled implicitly through faster degradation rates. Additionally, Ste7 is proposed to be hyperphosphorylated upon activating Fus3, tagging it for ubiquitination and subsequent degradation [149]. This is implemented with a degradation rule that is dependent on Ste7's phosphorylation state.

Ste5 also plays a role in regulating signal throughput, via feedback phosphorylation by Fus3 as well as by its presence in the cytoplasm. Ste5 is shuttled out of the nucleus, where it is normally sequestered, upon pheromone stimulation, though the precise mechanisms are unknown[150]. Therefore, we implement this export rate (S_{exp}) as an equation where G_f is the number of Gpa1's that are not bound to a Ste4:

$$S_{exp} = 0.3 \cdot \left(\frac{G_f}{G_f + 2500} \right). \tag{A.1}$$

This is the only rule that does not follow the law of mass action and was obtained from Shao *et al.*'s export rate equation. The only difference in this case is that our rate does not have a basal value of 0.0003 when $G_f = 0$.

Interaction or Reaction	Rate parameter(s)	Source	Cat.
Ste4/Ste20 interaction	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	0.8 s^{-1}	derived	I
Ste4/Ste5 assoc.	0.001725 molec^{-1}s^{-1}	derived	I
	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
Ste4/Ste5 dissoc.	0.2 s^{-1}	derived	I
	0.02 s^{-1}	derived	I
Ste5/Ste5 dimerization	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	0.001725 molec^{-1}s^{-1}	derived	I
Ste5/Ste5 dissoc.	0.075 s^{-1}	N/A	U
	0.0075 s^{-1}	N/A	U
	0.0005 s^{-1}	N/A	U
Ste11/Ste5 interaction	8.595×10^{-5} molec^{-1}s^{-1}	OM	D
	0.1605 s^{-1}	OM	D
Ste11 phosphorylation	0.5 s^{-1}	[36]	I
	0.5 s^{-1}	[36]	I
	0.5 s^{-1}	[36]	I
Ste5/Ste7 interaction	8.595×10^{-5} molec^{-1}s^{-1}	OM, [55]	D
	8.595×10^{-7} molec^{-1}s^{-1}	OM, [55]	I
	0.153 s^{-1}	OM, [55]	D
Ste7 phosphorylation	0.495 s^{-1} (12 rules)	OM	I (12)
MAPK/Ste7 interaction	4.35×10^{-6} molec^{-1}s^{-1}	OM	D
	0.0075 s^{-1}	OM	D
Fus3 phosphorylation	7.5 s^{-1} (4 rules)	OM	I (4)
Kss1 phosphorylation	1.5 s^{-1} (4 rules)	OM	I (4)
MAPK autophosphorylation	4×10^{-4} s^{-1}	N/A	U (2)
Ste5/Fus4 interaction	8.595×10^{-5} molec^{-1}s^{-1}	OM	D
	1.425 s^{-1}	OM	D

Table A.5: MAPK cascade interactions

Interaction or Reaction	Rate parameter(s)	Source	Cat.
Ste11 autodephos.	0.00087 (4 rules)	[36]	I (4)
Ste7 autodephos.	0.00087 (2 rules)	[36]	I (2)
Ste11/MAPK interaction	8.595×10^{-5} molec^{-1}s^{-1}	OM	I (2)
	1.5 s^{-1}	derived	I (2)
	0.75 s^{-1}	derived	I (2)
	0.75 s^{-1}	derived	I (2)
	0.375 s^{-1}	derived	I (2)
Ste11 phos. (by both MAPKs)	1.5 s^{-1}	OM	I (2)
Ste11 degradation	0.00075	OM	I
Ste5 phos.	1.5 s^{-1}	OM	I
Ste5 autodephos.	0.0087	[36]	I
MAPK autodephos.	0.00087 (4 rules)	[36]	I (4)
MAPK/Msg5 assoc.	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
Fus3/Msg5 dissoc.	7.5 s^{-1}	OM	I
	3 s^{-1}	OM	I
	3 s^{-1}	OM	I
	3 s^{-1}	OM	I
Kss1/Msg5 dissoc.	1.2 s^{-1}	OM	I
	0.12 s^{-1}	OM	I
	0.12 s^{-1}	OM	I
	0.12 s^{-1}	OM	I
MAPK dephos. (Msg5, both MAPKs)	0.12 s^{-1}	OM	I (2)
	0.12 s^{-1}	OM	I (2)
MAPK/Ptp assoc.	8.595×10^{-5} molec^{-1}s^{-1}	derived	I (2)
Fus3/Ptp dissoc.	1.5 s^{-1}	OM	I
	0.3 s^{-1}	OM	I
Kss1/Ptp dissoc.	0.15 s^{-1}	OM	I
	0.03 s^{-1}	OM	I
Fus3 dephos. by Ptp	1.2 s^{-1}	[36]	I
Kss1 dephos. by Ptp	0.12 s^{-1}	[36]	I

Table A.6: MAPK regulation interactions

Interaction or Reaction	Rate parameter(s)	Source	Cat.
Fus3 degradation	0.0002 s^{-1}	N/A	U
Msg5 degradation	0.0008 s^{-1}	N/A	U
Ste7 hyperphos.	0.495 s^{-1}	N/A	U
Ste7 degradation	0.002 s^{-1}	[36]	I
Ste(11,7)/Phosphatase interaction	7.155×10^{-5} $molec^{-1}s^{-1}$ 0.6 s^{-1}	N/A N/A	U (2) U (2)
Ste(11,7) dephos.	0.25 s^{-1} (5 rules)	N/A	U (5)
Ste5 nuclear import	0.5 s^{-1}	[36]	I

Table A.7: MAPK regulation interactions, continued

A.1.6 Nuclear interactions and regulation

Once Fus3 is activated, it translocates to the nucleus where it plays an active role in regulating genes associated with mating. Specifically, it inhibits Dig1 and Dig2 activity via phosphorylation. These two proteins, when not phosphorylated, bind to the transcription factor, Ste12, and prevent it from activating mating-related genes [35].

A.1.7 Gene interactions and protein synthesis

Upon pheromone stimulation, a number of genes in the mating cascade itself are expressed at higher levels (*STE2, SST2, GPA1, STE4, FUS3,* etc.), providing a measure of feedback [151]. Basal transcription of certain proteins is also present in the model.

A.1.8 Constructing the machine model

A number of rules were specifically designed to create a model that could assemble signaling machines. The interactions were arranged in a hierarchy to mimic the assembly of experimentally characterized machines [66]. Specifically, in the machine model Ste5 can only bind a Ste4 that is bound to a Ste20. In order to proceed to the MAPK cascade, a dimer of the Ste5-Ste4-Ste20 trimers must form. This hexamer can assemble in two ways: a trimer can bind another trimer,

Interaction or Reaction	Rate parameter(s)	Source	Cat.
Dig1/Ste12 interaction	8.595×10^{-4} molec^{-1}s^{-1}	derived	I
	30 s^{-1}	N/A	U
	3 s^{-1}	N/A	U
	3 s^{-1}	N/A	U
	0.003 s^{-1}	N/A	U
Dig2/Ste12 interaction	8.595×10^{-4} molec^{-1}s^{-1}	derived	I
	30 s^{-1}	OM	U
	3 s^{-1}	OM	U
Fus3/Ste12 interaction	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	3 s^{-1}	OM	U
	15 s^{-1}	OM	U
	15 s^{-1}	OM	U
	75 s^{-1}	OM	U
	0.3 s^{-1}	OM	U
	1.5 s^{-1}	OM	U
	1.5 s^{-1}	OM	U
	7.5 s^{-1}	OM	U
Kss1/Ste12 interaction	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	0.75 s^{-1}	OM	U
	3.75 s^{-1}	OM	U
	3.75 s^{-1}	OM	U
	18.75 s^{-1}	OM	U
	0.075 s^{-1}	OM	U
	0.375 s^{-1}	OM	U
	0.375 s^{-1}	OM	U
	1.5 s^{-1}	OM	U
Fus3/Dig1 interaction	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	4.5 s^{-1}	OM	I
	2.25 s^{-1}	OM	I
	2.25 s^{-1}	OM	I
	1.125 s^{-1}	OM	I

Table A.8: Nuclear interactions and regulation

Interaction or Reaction	**Rate parameter(s)**	**Source**	**Cat.**
Kss1/Dig1 interaction	8.595×10^{-6} molec^{-1}s^{-1}	derived	I
	7.5 s^{-1}	OM	I
	3.75 s^{-1}	OM	I
	3.75 s^{-1}	OM	I
	1.875 s^{-1}	OM	I
Fus3/Dig2 interaction	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	1.5 s^{-1}	OM	I
	0.75 s^{-1}	OM	I
	0.75 s^{-1}	OM	I
	0.375 s^{-1}	OM	I
Kss1/Dig2 interaction	8.595×10^{-5} molec^{-1}s^{-1}	derived	I
	2.55 s^{-1} OM	I	
	1.275 s^{-1} OM	I	
	1.275 s^{-1} OM	I	
	0.645 s^{-1}	OM	I
Dig phos. (2 MAPKs, 2 Dig proteins)	1.5 s^{-1}	OM	I (4)
Dig dephos. (2 Dig proteins)	0.00087 s^{-1}	[36]	I (2)
Dig2 degradation	0.0002 s^{-1}	N/A	U

Table A.9: Nuclear interactions and regulation, continued

Interaction or Reaction	Rate parameter(s)	Source	Cat.
Ste12/Ste2 gene interaction	2.145×10^{-5} molec^{-1}s^{-1}	derived	U
	0.03 s^{-1}	derived	U
Ste2 synthesis	3 s^{-1}	OM	I
	12 s^{-1}	[54]	D
Ste12/Gpa1 gene interaction	2.145×10^{-3} molec^{-1}s^{-1}	derived	U
	0.03 s^{-1}	derived	U
Gpa1 synthesis	27 s^{-1}	OM	I
Ste12/Ste4 gene interaction	2.145×10^{-4} molec^{-1}s^{-1}	derived	U
	0.03 s^{-1}	derived	U
G-protein basal synthesis	0.5 s^{-1}	OM	I
Ste4 synthesis	18 s^{-1}	OM	I
Ste12/Sst2 gene interaction	2.145×10^{-5} molec^{-1}s^{-1}	derived	U
	0.03 s^{-1}	derived	U
Sst2 synthesis	0.78 s^{-1}	OM	I
	1.5 s^{-1}	OM	I
Ste12/Fus3 gene interaction	2.145×10^{-5} molec^{-1}s^{-1}	derived	U
	0.03 s^{-1}	derived	U
Fus3 synthesis	4 s^{-1}	OM	I
	15 s^{-1}	OM	I
Ste12/Msg5 gene interaction	2.145×10^{-5} molec^{-1}s^{-1}	derived	U
	0.03 s^{-1}	derived	U
Msg5 synthesis	0.08 s^{-1}	OM	I
	0.63 s^{-1}	OM	I
Ste12/Dig2 gene interaction	2.145×10^{-5} molec^{-1}s^{-1}	derived	U
	0.03 s^{-1}	derived	U
Dig2 synthesis	0.24 s^{-1}	OM	I
	0.45 s^{-1}	OM	I
Ste12/Ste12 gene interaction	2.145×10^{-5} molec^{-1}s^{-1}	derived	U
	0.03 s^{-1}	derived	U
Ste12 synthesis	0.45 s^{-1}	OM	I,I

Table A.10: Gene interactions and protein synthesis

or a trimer can bind a free Ste5 and subsequently bind a Ste4-Ste20 dimer. Then Ste11 can bind Ste5, and in order for Ste7 to bind Ste5, both Ste5 proteins must be bound to a Ste11 (forming an octomer). Only after both Ste7 proteins have bound to assemble the full decamer structure can phosphorylation occur. Once all four kinases are fully phosphorylated, the machine binds and phosphorylates Fus3 as a multi-subunit kinase.

Dissociation of the machine can proceed in two ways. The first is a disassembly pathway which is essentially the inverse of the assembly pathway. However, once the decamer is fully assembled, our rates are implemented in a way that reflects the inherent stability of a machine's quaternary structure [66]. Thus we adapted the Ste7 hyperphosphorylation mechanism (that is present in the ensemble model, Section A.1.5) to promote rapid dissociation of the signaling machine into its constituent monomers, and mimic the ability of Fus3 to induce negative feedback [36].

The following table of rates are those involved with rules that mechanistically differ from the ensemble model. Since these interactions were invented in the absence of any experimental evidence, the rates, in addition to the mechanisms, are hypothetical and were implemented in order to replicate experimental time-course and dose-response trends. The association rates were designed to be as similar as possible to those present in the ensemble model, however a few required manipulation to match experimental data. Varying the equation governing Ste5 nuclear export (S_{exp}^{mach}) was the primary means of replicating the dose-response curve. In particular it was altered to be more sensitive to the amount of free Gpa1 in the form of a Hill function:

$$S_{exp}^{mach} = 0.3 \cdot \left(\frac{{G_f}^4}{{G_f}^4 + 12000^4} \right). \tag{A.2}$$

The subsequent table are those reaction events that have identical mechanisms but different rate constants between the two models.

Novel machine model events

Interaction or Reaction	Rate parameter(s)	Source	Cat.
Ste5/Ste4-20 interaction	8.595×10^{-5} molec^{-1}s^{-1}	N/A	U
	0.2 s^{-1}	N/A	U
Ste5-4-20/Ste5 interaction	8.595×10^{-5} molec^{-1}s^{-1}	N/A	U
	0.2 s^{-1}	N/A	U
Ste5-4-20-5/Ste4-20 interaction	0.001725 molec^{-1}s^{-1}	N/A	U
	0.0005 s^{-1}	N/A	U
Ste5-4-20/Ste5-4-20 interaction	8.625×10^{-4} molec^{-1}s^{-1}	N/A	U
	0.0005 s^{-1}	N/A	U
hexamer/Ste11 interaction	8.595×10^{-4} molec^{-1}s^{-1}	N/A	U
	0.01605 s^{-1}	N/A	U
octomer/Ste7 interaction	8.595×10^{-4} molec^{-1}s^{-1}	N/A	U
	8.595×10^{-6} molec^{-1}s^{-1}	N/A	U
	0.0153 s^{-1}	N/A	U
	1.53×10^{-6} s^{-1}	N/A	U
decamer activation (phos.)	0.1 s^{-1} (5 rules)	N/A	U
MAPK/decamer interaction	8.595×10^{-5} molec^{-1}s^{-1}	N/A	U
	0.075 s^{-1}	N/A	U
MAPK dissoc.	10 s^{-1} (2 MAPKs)	N/A	U (2)
MAPK phos.	0.1 s^{-1} (2 rules, 2 MAPKs)	OM	I (2,2)
Ste7 hyperphos.	8.595×10^{-5} molec^{-1}s^{-1}	N/A	U
	0.1 s^{-1}	N/A	U
	0.1 s^{-1}	N/A	U
	1 s^{-1}	N/A	U
Ste7 dephos. (alternate site)	0.0087	[36]	I

Table A.11: Novel machine model events

Identical reactions with different rates

Interaction or Reaction	Ensemble model rate	Machine model rate
Sst2/Ste2 dissociation	0.15 s^{-1}	0.015 s^{-1}
Ste4/Ste20 interaction	8.595×10^{-5} $molec^{-1}s^{-1}$ 0.8 s^{-1}	8.595×10^{-4} $molec^{-1}s^{-1}$ 0.08 s^{-1}
Ste5/Fus3 dissoc.	1.425 s^{-1}	1 s^{-1}
Ste11/7 autodephos.	0.00087 s^{-1} (6 rules)	0.0087 s^{-1} (6 rules)
Msg5/Kss1 dissoc.	1.2 s^{-1} 0.12 s^{-1} (3 rules)	0.12 s^{-1} 0.012 s^{-1} (3 rules)
Fus3/Dig1 assoc.	8.595×10^{-5} $molec^{-1}s^{-1}$	8.595×10^{-4} $molec^{-1}s^{-1}$
Kss1/Dig1 assoc.	8.595×10^{-6} $molec^{-1}s^{-1}$	8.595×10^{-5} $molec^{-1}s^{-1}$
Dig phos. (2 MAPK, 2 Dig proteins)	1.5 s^{-1} (4 rules)	15 s^{-1} (4 rules)
Dig dephos. (2 Dig proteins)	0.00087 s^{-1} (2 rules)	0.001 s^{-1} (2 rules)
Ste5 phos.	1.5 s^{-1}	1 s^{-1}
Ste(11,7) dephos.	0.25 s^{-1} (5 rules)	1 s^{-1} (5 rules)

Table A.12: Identical reactions with different machine model rates

A.2 Model Simulation

Our model was simulated using rule-based techniques, in particular the Kappa rule-based modeling language [15, 39, 58, 145, 146] and its associated simulator, KaSim [16]. As mentioned in the main text, these methods allow us to incorporate and investigate the influence of combinatorial complexity in our modeling without needing to explicitly enumerate all the possible species that our model can generate [12, 16]. This section describes how we implemented and simulated our model as well as our method for randomizing our parameters in order to characterize the robustness of drift (see Section A.3.3) in our simulations.

A.2.1 Kappa and KaSim

Kappa and related languages employ rules to define reaction network dynamics. Stochastic simulation of these rules using KaSim involves an adapted version of the Gillespie algorithm [89]. Briefly, a rule has a left hand side (LHS), a right hand side (RHS) and an associated stochastic rate constant. Both the LHS and RHS are site graphs represented by Kappa strings. These are particular patterns (before and after the reaction event) that could be present in a simulation's *mixture* of explicitly represented protein agents (which form the set of complexes at any given time) [145, 146]. At each event, a particular rule is selected with a probability proportional to the number of LHS pattern matches in a mixture and the rule's rate constant. During simulation, KaSim can output specified observables (typically the number of matches of a Kappa string) at uniform increments of time. In addition to this time-course data we also employed the snapshot mechanism. This outputs the entire mixture (all present species) at a specified point in time. Both snapshots and observables were used in our analyses of the models.

For more information on the Kappa language itself, see [145, 146]; descriptions of the simulation algorithm implemented in KaSim are present in [12, 16]. Finally, KaSim itself is open-source software and can be downloaded at `http://github.com/jkrivine/KaSim`.

A.2.2 BioNetGen and NFsim

In order to confirm that our results were consistent with other rule-based modeling methods, we implemented the machine and ensemble model in the BioNetGen language [14]. We see identical mean trajectories for NFsim and KaSim in Figs. A.1 and A.2 (additional comparisons can be seen in Section A.4 and Figs. A.4 and A.5). We also employed the network generation tool present in BioNetGen to enumerate all the possible species that the machine model can create; this process failed for the ensemble model due to memory restrictions (see Section A.3.5).

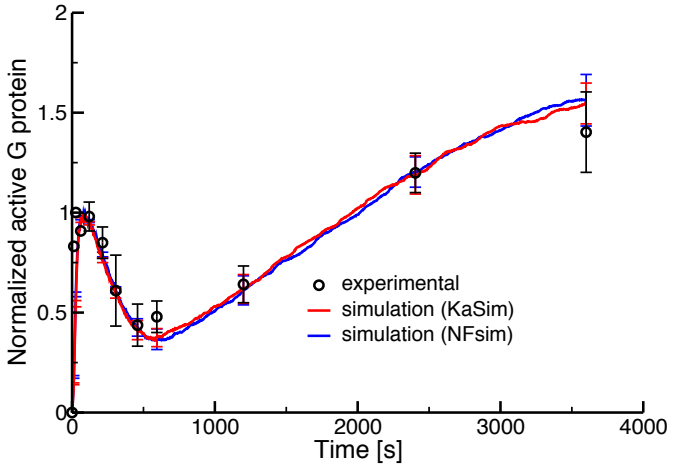


Figure A.1: Comparison of G protein activation dynamics using NFsim and KaSim

A.2.3 Simulation methods

We employ a specific method for the simulation of our model in order to produce the most realistic results possible. Our method is outlined in Fig. A.3. Since actual cells do not contain sets of monomers, we perform simulations in the absence of pheromone to generate an unstimulated *steady-state*. Specifically, we simulate our model for 1000 seconds starting from initial conditions which involve all agents in their monomeric state, aside from Gpa1 and Ste4, which are found in Gpa1/Ste4 heterodimers. Upon completion of N simulations, we output the mixtures as

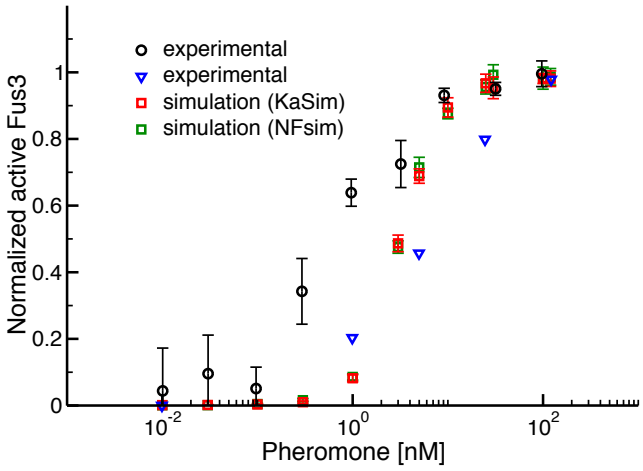


Figure A.2: Comparison of dose-response trends using NFsim and KaSim

snapshots and use these sets of complexes as new sets of initial conditions. This simulated set of steady-states can be considered as representative of a population of N untreated yeast cells. For each steady-state, we add pheromone to induce the mating response (100 nM in all cases except the dose-response simulations) and then generate N' hour-long (in simulated time) trajectories, resulting in $N \times N'$ total signaling simulations.

For the drift calculations (see Section A.3.3) we output snapshots on a logarithmic time-scale, and execute our pairwise comparisons between all simulations that originated from the same steady-state simulation (*e.g.* Fig. 2.3A in the main text has $N = 1$ and $N' = 10$ resulting in $\binom{N'}{2} = 45$ unique pairwise comparisons for each time point). This allows us to observe the heterogeneity among complexes generated solely from signaling (addition of pheromone), rather than from differences present before pheromone stimulation.

A.2.4 Parameter randomization

To examine the robustness of drift with respect to the chosen parameters, we generated 1000 different parameter sets for the ensemble and machine models in a manner similar to prior studies

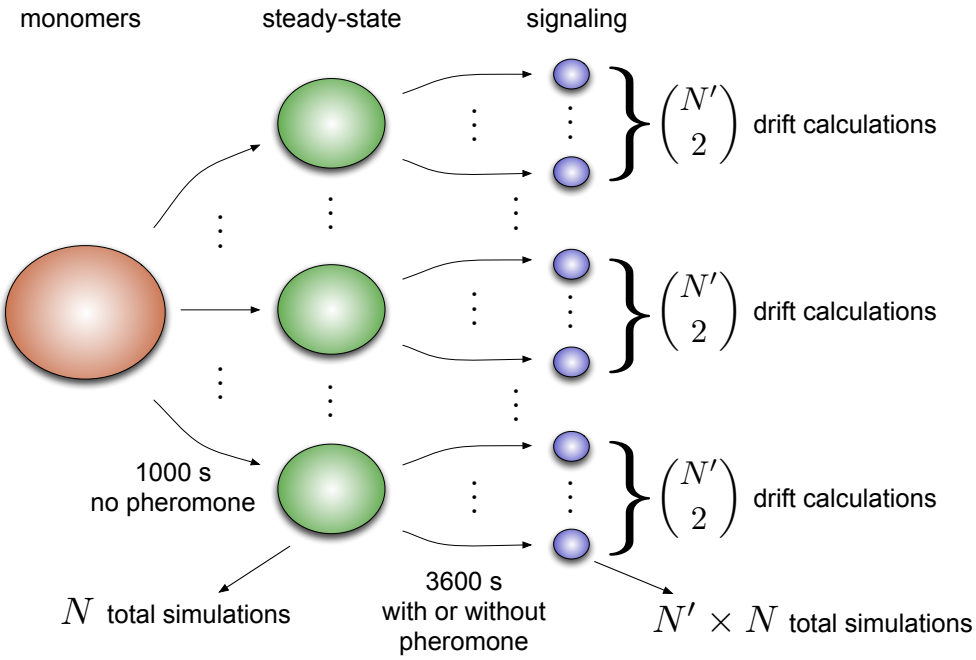


Figure A.3: General method for calculating drift. A rule set is initialized with a specific set of proteins (red) and simulated for 1000 seconds to N independent steady-states (green). The steady-states are then used to generate N' signaling simulations each (blue, 1 hour of simulated time each), resulting in a total of $N \times N'$ simulations from the steady-state conditions for a particular rule or parameter set. Pairwise drift calculations are only performed between simulations with identical initial conditions, thus for each steady-state we have $\binom{N'}{2}$ drift values resulting in a total of $N \times \binom{N'}{2}$ drift points for any specific rule set. Note that we also performed the "signaling" simulations *without* pheromone to observe the baseline levels of drift in our model (seen in the main text, Fig. 2.3A).

[44]. For all non-varied and inferred or estimated rate parameters, the particular value was multiplied by a uniformly sampled number, x: $10^{-1} < x < 10^{1}$. For those parameters varied (seen above in red, Section A.1.2) as well as those directly observed, the rate was also multiplied by a uniformly sampled number, but on a smaller range: $2^{-1} < x < 2^{1}$. This was done to maintain a level of realism in these randomizations, as the varied parameters typically have more influence over the experimentally determined time-course trends. Despite this, there was still noticeable deviation from wild-type behavior in these simulations (both time-course and dose-response) due to

the wide range of parameter variation.

Upon generation of these parameter sets, each was simulated to $N = 3$ unique steady-states (Fig. A.3). Subsequently, pheromone was added and $N' = 3$ trajectories were simulated, resulting in nine simulations of the signaling network for each unique parameter set. Thus for each set, we have three sets of three drift values, resulting in 9000 total drift points for the ensemble model, and slightly fewer (7789) for the machine model as simulation pairs where $d(i, j) = 0$ were ignored; these cases represented parameter sets that had essentially no signaling activity. Figs. 3b and 5b in the main text show this density distribution alongside distributions of drift values from the validated ensemble and machine parameter sets. In order to generate these distributions we used 50 copies of the final (ensemble or machine) model instead of 1000 unique parameter sets, resulting in 450 drift points (again, $N = 3$ and $N' = 3$). These densities were generated using kernel density estimation (KDE) in the R statistical suite [113]. We determined significant differences between the means of such distributions via a non-parametric two-tailed permutation test with 10^5 replications (also in R [113]).

A.3 Additional results

A.3.1 Model validation

Since each simulation requires approximately 3-4 hours of CPU time, standard methods of fitting our model to data (*e.g.* regression techniques) could not be implemented [62, 152]. We thus manually varied parameters in the model in order to match experimental data. To do this, we identified a subset of parameters that govern experimentally observed trends [37, 38, 54, 153]. We modified these parameters iteratively, within a biologically realistic range (see Section A.1.2), to achieve reasonable overlap with experimental observation. It is important to note that the model reproduces these observations employing mass-action kinetics and does not utilize simplified Michaelis-Menten functions [36]. Though the model reproduces the time- and pheromone-dependent trends of the yeast pheromone response cascade, it is clearly not the only possible solution, since we

were able to construct a machine-like model which also replicates certain experimental trends (see Section A.3.2).

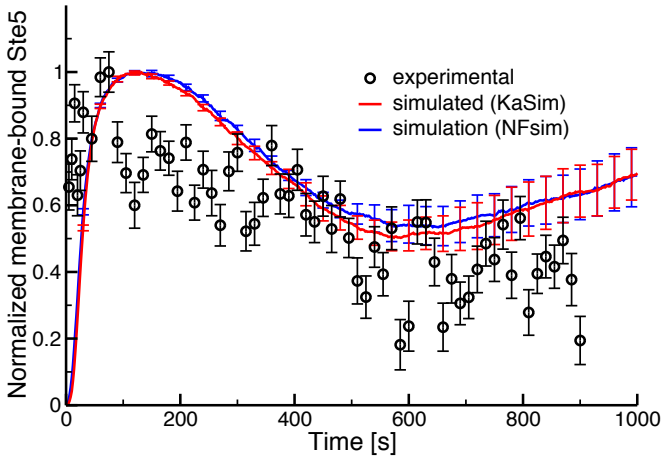


Figure A.4: Activation of the pheromone cascade in the ensemble model results in rapid localization of the scaffold, Ste5, to the membrane as indicated by FRET measurements [37]. Values are seen for the first 1000 seconds and the error bars represent 95% confidence intervals for both experimental and simulated data ($n = 3$ and $n = 10$, respectively).

The two graphs seen here (Fig. A.4 and Fig. A.5), in addition to those in the main text (Fig. 2.2), are the experimental trends [37, 153] that were used to validate our model. Our data is broadly consistent with experimental data (*e.g.* the initial spike in Ste5 membrane-recruitment in Fig. A.4), especially when considering the noise present in the experimental measurements and the potential impacts of photobleaching [37]. Note that of the four sets of experimental data (including those described in the main text), two are directly affected by the concentration of unbound Ste4 (*i.e.* not bound to Gpa1): Ste4-Ste20 binding and active G-protein. Thus among the most influential unknown and varied parameters were the Ste12-induced synthesis rates of Ste4 and Gpa1.

A.3.2 Machine model validation

Validation of the machine model was accomplished in essentially the same way as validation of the ensemble model. Only very minor adjustments to the rates were needed to reproduce the G-

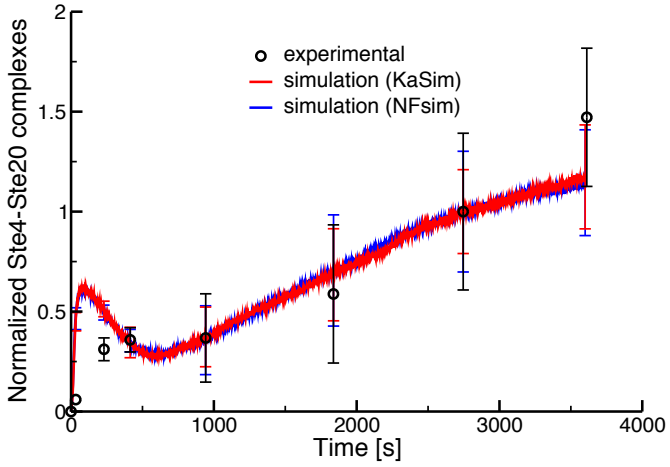


Figure A.5: Fold increase over the basal number of Ste4-Ste20 dimers in the ensemble model [153]. The error bars represent 95% confidence intervals for both experimental and simulated data ($n = 3$ and $n = 10$, respectively).

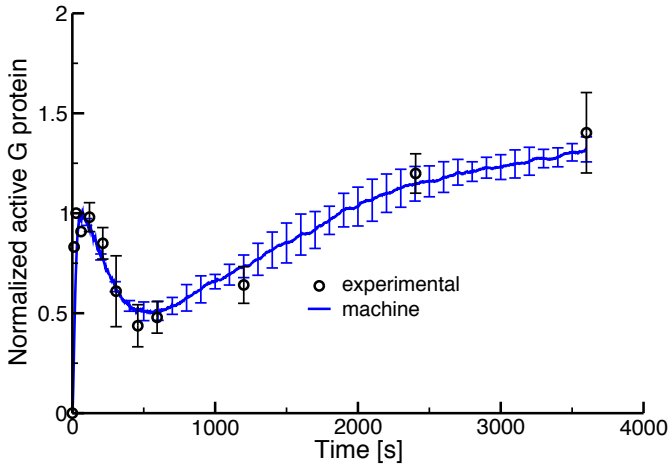


Figure A.6: G-protein activation dynamics in the machine model [54]. Error bars are 95% confidence intervals, experimental data is seen in black ($n = 3$), and simulations are seen in blue ($n = 10$).

protein temporal dynamics (Fig. A.6), and the dose-response curve was readily matched upon altering the cooperativity of the Ste5 nuclear export rate (Section A.1.8, Fig. A.7). Though not as accurate as the ensemble model, a similar trend for Ste4-Ste20 binding was seen in the machine model (Fig. A.8). However, it was unable to exactly reproduce the behavior seen in Fig. A.4.

This was most likely due to the altered rules, which require Ste5 and its binding partners to remain membrane-bound in order to phosphorylate Fus3.

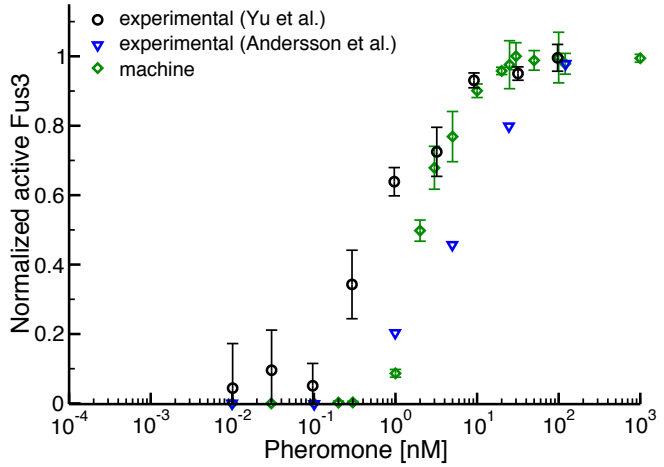


Figure A.7: Dose-response dynamics in the machine model (phosphorylated Fus3 with respect to pheromone) [37, 38]. Error bars are 95% confidence intervals. Data from [37] ($n = 3$) and [38] ($n =$ unknown) are in black and blue, respectively. Simulated data is in green ($n = 10$).

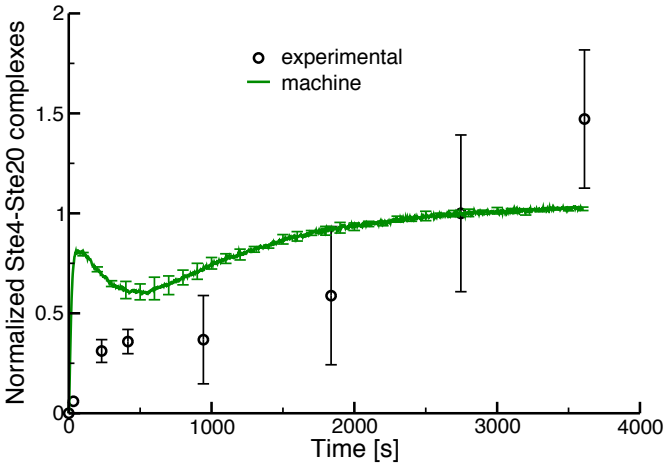


Figure A.8: Fold increase over the basal number of Ste4-Ste20 dimers in the machine model. [153] The error bars represent 95% confidence intervals for both experimental and simulated data ($n = 3$ and $n = 10$, respectively).

A.3.3 Compositional drift

Compositional drift was first introduced as a measure of intracellular heterogeneity in [12]. Drift (d) is a pairwise comparison between the set of complexes, C, of two independent simulations, i and j, which originated from the same initial conditions. It is defined as the symmetric difference of the two sets divided by the union of the two sets:

$$d(i,j) = \frac{|C_i \Delta C_j|}{|C_i \cup C_j|} \tag{A.3}$$

where $|X|$ is the number of elements in some set X. This results in a normalized value between 0 and 1 where $d = 0$ indicates identical sets of complexes and $d = 1$ indicates disjoint sets. Given two simulated cells and their constituent complexes, drift is thus the probability that a given complex is present in one cell but not the other.

As mentioned in the main text, this calculation takes into account any difference between two complexes, however minor. To confirm that this is a reasonable method of determining the level of heterogeneity among signaling species, we examined a number of different criteria. First, we took snapshots from ten simulations and calculated drift while ignoring any difference due to post-translational modifications (*e.g.* phosphorylation). We can see that although drift is reduced slightly in this case, substantial heterogeneity still exists when solely considering the binding patterns of the present complexes (Fig. A.9).

However, phosphorylation is certainly important in a signaling cascade; Fus3 cannot induce transcription of mating-related genes without being phosphorylated on two of its residues. Thus we investigated a comparison that can incorporate these distinctions while ignoring differences among complexes which have no behavioral consequence for the system. We used our ten snapshots to construct classes of complexes that are *functionally equivalent*. For two complexes to be functionally equivalent, the system (or rule set in our case) must not be able to distinguish between the complexes. In essence, one could exchange these two complexes without having any effect on the behavior at that point in time. We defined these equivalence classes using our rule set. In order

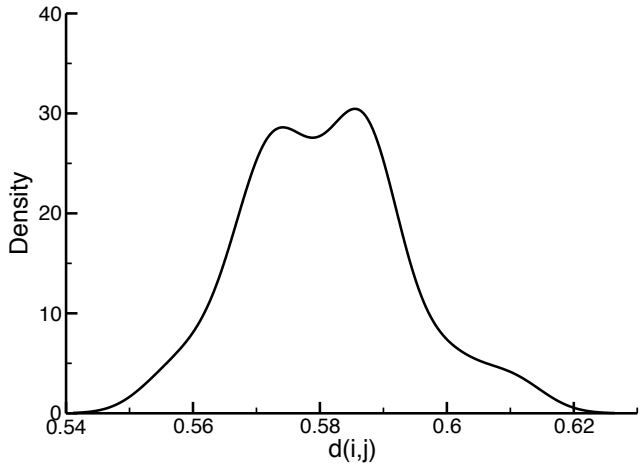


Figure A.9: Drift density in the ensemble model at $t = 360$ seconds without consideration of post-translational modification ($n = 45$). The density was estimated using standard KDE methods in R [113]

to determine which reaction should take place next, each rule is assigned a stochastic rate of being chosen, called the *activity* [16]. The activity is the product of the rate constant and the number of complexes in the mixture that match the LHS of the rule. The number of matches is important for a specific reason: if we have a rule where A converts to B at a rate k and there is no A in the mixture, the system cannot execute this rule regardless of its rate, thus the rule's activity is 0. By calculating the activity of every rule with respect to a particular complex, we obtain a signature for this complex within a particular rule set. If two (or more) complexes exhibit the same signature (*i.e.* every rule has the same activity) then they belong to the same equivalence class, because they exert the same influence on the system. We found that no two complexes were functionally equivalent over a set of 10 simulations, indicating that the structural distinctions included in our original definition of drift are functionally relevant [12].

In addition to examining the level of drift during peak signaling (Figs. 3b and 5b in the main text) we also examined drift over the union of logarithmically distributed time points. To do this, we compiled a list of the observed unique species from all time points for a particular simulation and performed the pairwise drift calculation with a similarly compiled list of unique species from another simulation. We found very similar results when calculating the density for $N = 50$

simulations (450 total drift values; Fig. A.10), as compared to the drift densities at the single peak signaling time point ($t = 360$). This confirms that the large differences in generated species between simulations is not an artifact of choosing a single time point for the calculation.

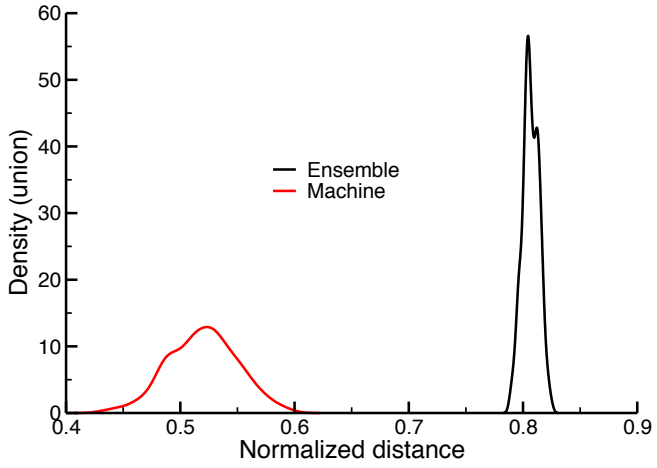


Figure A.10: Drift density for scaffold-based signaling species over multiple, logarithmically distributed time points. Similar to Figs. 3B and 5B in the main text, we see a stark difference between the machine and ensemble models' average drift. Clearly the drift between simulations is not a result of the time point for which the drift calculation is made.

We also investigated the rate at which a particular simulation diverged from its initial conditions based on drift. This was termed "autodrift," and is defined as the drift between a simulated cell i's sets of complexes at two different points in time: $d_i(t, t+\Delta t)$. We fit the data to an exponential function using standard nonlinear least-squares regression in R [113]. Analysis of the residuals indicated that a single exponential fit did not capture the trend in the data. We therefore attempted fits using both double and triple exponential functions. The functional form of the full model is:

$$d_i(t, t+\Delta t) = \beta_1 - \beta_2 \cdot e^{-\beta_3 \Delta t} - \beta_4 \cdot e^{-\beta_5 \Delta t} - \beta_6 \cdot e^{-\beta_7 \Delta t}. \tag{A.4}$$

We found that fitting the entire model yielded an estimate for the third exponential term (*i.e.* $\hat{\beta}_7$) that was not statistically significant when correcting for multiple hypothesis testing. All the other

Parameters	Model Estimates (p-value)		
	single	**double**	**triple**
$\hat{\beta}_1$	0.7633 ($< 2 \times 10^{-16}$)	0.7759 ($< 2 \times 10^{-16}$)	0.7850 ($< 2 \times 10^{-16}$)
$\hat{\beta}_2$	0.6821 ($< 2 \times 10^{-16}$)	0.2378 ($< 2 \times 10^{-16}$)	0.2693 ($< 2 \times 10^{-16}$)
$\hat{\beta}_3$	5.452 ($< 2 \times 10^{-16}$)	2.090 ($< 2 \times 10^{-16}$)	2.609 ($< 2 \times 10^{-16}$)
$\hat{\beta}_4$	N/A	0.5252 ($< 2 \times 10^{-16}$)	0.4849 ($< 2 \times 10^{-16}$)
$\hat{\beta}_5$	N/A	10.82 ($< 2 \times 10^{-16}$)	11.65 ($< 2 \times 10^{-16}$)
$\hat{\beta}_6$	N/A	N/A	0.02013 (4.98×10^{-16})
$\hat{\beta}_7$	N/A	N/A	0.01795 (0.005)

Table A.13: Autodrift fitting parameters

estimates in all of the variants of the model were significant (see table below). Among this set of nested models we thus selected the double exponential fit as the one that most significantly describes the data. Close inspection of the residuals indicates that there may indeed be a significant further increase in drift on longer time scales (Fig. 3c, main text); the difficulty in this case is that the time window is not long enough to capture this trend in a significant way. Longer simulations could thus yield a model where $\hat{\beta}_7$ is more significant. It is likely that these longer timescale changes in the value of drift represent more than just the turnover of transient complexes, but rather susbstantive changes in the system that arise due to the progress of the signal down the cascade.

A.3.4 Species classification and clustering

Since most of the combinatorial complexity in this cascade is centered around Ste5 and all its potential interaction partners, we classified the complexes present in our snapshots into 6 categories or *bins*. In order to do this we constructed the largest possible complexes starting from monomers based on the rule set. The resulting 6 complexes (with Fus3 and Kss1 considered interchangeable) are the basis for classification; any species generated during simulation will match a pattern present in one of these complexes and thus belong to its bin.

Three of the six bins are directly related to specific aspects of the response network: one bin

contains all G-protein related complexes (and the pheromone peptide), another contains all species localized in the nucleus, and a third contains all scaffold-based species (note that some monomeric species could be placed in multiple bins: we placed these in the scaffold-based bin by convention). The remaining three bins are primarily composed of a kinase (Ste11, Ste7 and Fus3 and Kss1) and its specific phosphatase. Here we focus almost exclusively on the scaffold bin. A representation of the complex defining this bin is seen in Fig. A.11.

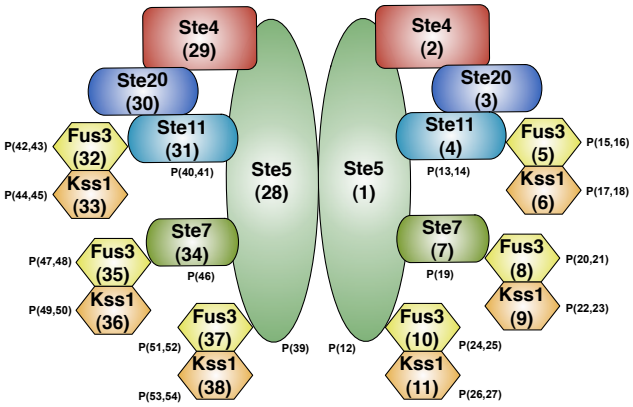


Figure A.11: This diagram shows a labeling system for creating the integer sequences for the clustering of complexes (take note that this complex can never actually exist in our simulations since Fus3 and Kss1 cannot simultaneously bind their shared substrates; it is merely a visual representation of how we create unique identifiers for each complex). Each position in the sequence is associated either with a protein agent or an agent's phosphorylation site. Note that some agents have multiple phosphorylation sites (*e.g.* in our model, Fus3 can be phosphorylated on two independent residues). Each number in the sequence can have a range of integer values. For protein agents this is either 0 or 1, indicating their presence in the complex. The range of potential integers for phosphorylation sites vary between sites, however 0 indicates no phosphorylation for all sites. Sites representing a specific residue are either 1 or 0 (indicating the presence/absence of phosphorylation), however sites representing multiple residues can have values larger than this (*e.g.* Ste11 has a site representing 3 distinct residues requiring phosphorylation for its activation, therefore this site can contain values up to 3).

Following our analysis of subgraph conservation among scaffold-based signaling species, we performed clustering analysis on this subset of complexes during peak signaling ($t = 360$ seconds). This enabled us to formalize our search for the existence of a core complex. In order to cluster the complexes generated by our simulations we converted each complex in a particular snapshot to a unique sequence of integers (vector) in order to calculate the *graph edit distance* (G_{edit}) between

any two species in the same bin. In graph theory, G_{edit} is the minimum number of *edits* or changes necessary to convert one graph to another. In the case of our depictions of macromolecular complexes, bonds are equivalent to edges (simple contact, with Ste20/Ste11 contact an exception) and protein or gene agents represent nodes. When considering these complexes in our vector notation, G_{edit} is the sum of the absolute value of differences at each position in two sequences, r and s, of length, l:

$$G_{edit} = \sum_{n=0}^{l} |r_n - s_n|. \tag{A.5}$$

This is also known as the "Manhattan distance", or L^1-norm. We included differences in phosphorylation states in G_{edit}, as these differences clearly have an effect on the signaling network (Section A.3.3). In Fig. A.11, we outline our method for creating these sequences. Note the symmetry in this scaffold-based bin, resulting in two possible integer sequences per complex and thus two unique ways to calculate G_{edit} between two particular species (as mentioned above, G_{edit} is always the minimum value in this situation). A sample calculation is seen in Fig. A.12.

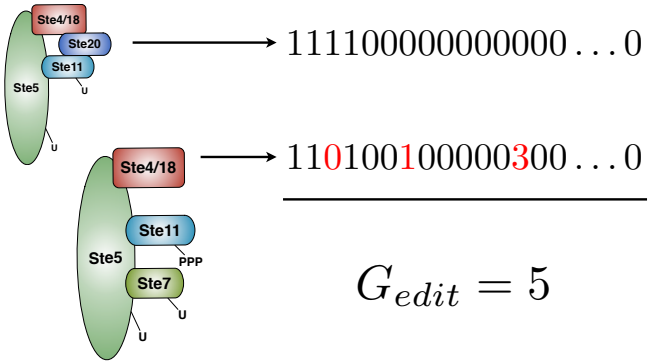


Figure A.12: Calculating G_{edit} between two complexes. Each sequence is one of two potential vector representations for its complex. In terms of *graph edits*, we can see that there are five (seen in red): removal of Ste20, addition of unphosphorylated Ste7, and addition of three phosphates on Ste11.

Upon generation of these sequences, we proceeded to hierarchically cluster the complexes according to the G_{edit} matrix. We focused on clustroid-based single-linkage clustering for our data though both standard single- and complete-linkage criteria gave similar results. We also attempted to find the optimal number of clusters, N, via a specific stopping criterion, $E(i)$ [154]. Unfortu-

nately this criterion did not return consistent results between multiple snapshots, thus we chose $N = 10$ as our ultimate number of clusters for analysis. Varying this cutoff did not influence the results discussed below (seen below and in Fig. 2.4B in the main text).

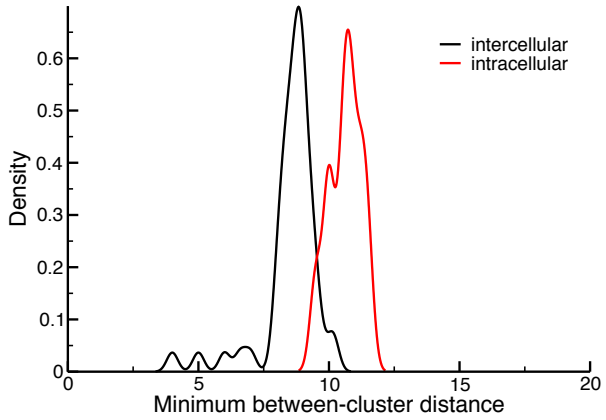


Figure A.13: MBCD distributions for clusters in the ensemble model both between and within snapshots ($n_{intra} = 90$ and $n_{inter} = 45$). Though the mean intercellular MBCD is lower than the mean intracellular MBCD ($p < 10^{-5}$) we would expect that the intercellular MBCD would be near 0 if there was conservation within the scaffold species. Both densities were estimated using KDE methods in R [113]

If the complexes expressed some sort of conserved binding pattern, we would expect that the clusters generated from one snapshot would be near identical or similar to the clusters generated from another. However this was not the case, as the intercellular minimum between-cluster distance (MBCD) is nearly as large as the intracellular MBCD (when hierarchically clustering the complexes, the intercellular MBCD is the criteria for selecting which two clusters will join next, Fig. A.13).

We also found that clusters containing more than 10 complexes exhibited very little conservation in terms of structural similarity between their constituent species. Note that in many of these clusters not even Ste5 dimers were conserved. In fact, the mean of the average G_{edit} between the clustroid and its cluster constituents (where the number of constituents ≥ 10) is greater than 6 as seen in Fig. A.14. A different way of framing this result is to consider the largest conserved component within a particular cluster (while still retaining the condition on the number of constituents). We calculate this conserved component using exclusively those entries in the vector

representation that refer to protein presence/absence; we do not consider phosphorylation state. With our standard clustering cutoff of $N = 10$ (Fig. A.15, black) we see a distribution where a conserved component consisting of only 2 proteins is in the 89^{th} percentile. As we increase N, the mean of this distribution changes slightly, but maintains consistently low values (near or around 2). We also see a shift in the peak of the distribution (from 0 to 2) between $N = 10$ and $N = 20$, however it remains at 2 through $N = 100$ indicating consistently low conservation within clusters of reasonable size. Thus we reach the same conclusion with our clustering that we did with our simpler analysis of structural conservation seen in Fig. 2.4A in the main text.

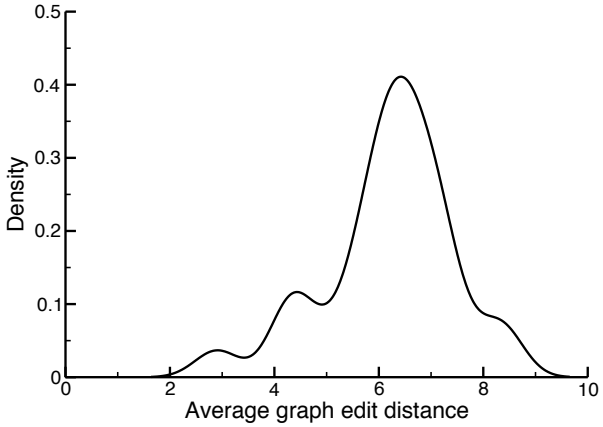


Figure A.14: Distribution of average G_{edit} scores between a clustroid and its constituent complexes in the ensemble model ($N = 26$). Note this is only calculated when the number of elements in a cluster is ≥ 10. Upon consideration of 10 independent simulations (and their snapshots) only 26 of the total 100 clusters contained over 10 complexes. Density estimated using KDE methods in R [113]

A.3.5 Enumerating all possible species

In order to determine the total number of scaffold-based (*i.e.* Ste5-bound) protein complexes that can possibly form, we proceeded to use BioNetGen to construct the reaction network for the two models. The output listed all species that could be formed from that particular ruleset, and from this list we found a total of 1106 Ste5-bound species for the machine model. BioNetGen was incapable of enumerating the species present in the ensemble model due to memory restrictions, so

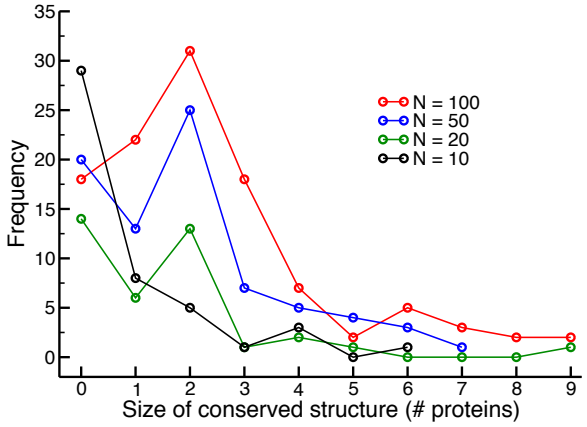


Figure A.15: Frequency of the size of the largest conserved component in clusters with 10 or more complexes. Different colors represent different distributions based on varying values of the clustering cutoff, N. Total number of clusters for each distribution: red = 110, blue = 78, green = 38, black = 47.

we analytically calculated the total number of scaffold-based species. We implemented a counting procedure made relatively simple by the mutual independence of many of the binding interactions in the ensemble model (*i.e.* Ste5 can bind Ste11 independent of all of Ste5's other binding sites and independent of Ste11's phosphorylation state). Consider the structure seen in Fig. A.11. Since all of Ste5's binding interactions are mutually independent we can focus on each of Ste5's sites individually and then take the product across all sites to estimate the total number of possible states. Initially we will focus on molecules that have *only one* Ste5.

We see that Ste5's top site binds Ste4 and Ste4 can bind Ste20. This results in 3 total species, assuming Ste5 is always present: Ste5, Ste5-Ste4, and Ste5-Ste4-Ste20, since Ste20 can only be present in the presence of Ste4. The second site on Ste5 binds Ste11 (again, independently of all other sites), and Ste11 can be in 4 phosphorylation states according to the agent declaration present in the Kappa model file:

%agent: Ste11(mapk, ste5, degradation~u~p, S302_S306_S307~u~p~pp~ppp)

Since we are primarily concerned with those structures directly related to signal transduction,

we will ignore the 'degradation' site for this calculation. Thus we have 5 different states (Ste5 unbound + 4 Ste5-Ste11 states). Fig. A.11 also shows the possibility of Fus3 *or* Kss1 binding to Ste11, and this interaction is independent both from Ste11's phosphorylation state and Fus3/Kss1's phosphorylation state. Since both Fus3 and Kss1 can be in 4 phosphorylation states, we have a total of 9 possible ways that Ste11 can bind a MAPK while bound to Ste5 (Ste5-Ste11 unbound + 4 Ste5-Ste11-Fus3 states + 4 Ste5-Ste11-Kss1 states). Finally, we multiply the 4 Ste5-Ste11 states times the 9 Ste5-Ste11-MAPK states and add the remaining unbound Ste5, resulting in 37 total states for Ste5's Ste11 binding site.

Further counting is as follows:

Ste5 site	**States**	**Description**
Ste4	3	Ste5 + Ste5-Ste4 + Ste5-Ste4-Ste20
Ste11	37	4 Ste11 states · 9 MAPK states + 1 unbound Ste5
Ste7	28	3 Ste7 states · 9 MAPK states + 1 unbound Ste5
MAPK	9	8 MAPK states + 1 unbound Ste5
phosphorylation	2	2 phosphorylation states
Total	55944	

From here we can then enumerate all species that include a Ste5 molecule. Since both Ste5 molecules can operate independently we have 55944^2 species with Ste5 dimers + 55944 species with a Ste5 monomer, resulting in nearly 3.13 billion Ste5-based species, nearly a 3 million-fold increase compared to the machine model.

A.3.6 Socio-affinity scoring

We used the socio-affinity (SA) index described in [19] to determine whether standard methods of deriving complex information from high-throughput data, such as tandem affinity purification (TAP), can distinguish between machine-like complexes and ensembles of signaling species. Note that the following definitions and equations merely summarize descriptions given in [19]. The SA

score between protein i and j, $A(i,j)$, is based on binary (bait and prey) interaction data, and is a linear combination of a number of terms:

$$A(i,j) = S_{i,j|i=bait} + S_{i,j|j=bait} + M_{i,j} \tag{A.6}$$

where

$$S_{i,j|i=bait} = \log\left(\frac{n_{i,j|i=bait}}{f_i^{bait} \cdot n_{bait} \cdot f_j^{prey} \cdot n_{i=bait}^{prey}}\right) \tag{A.7}$$

and

$$M_{i,j} = \log\left(\frac{n_{i,j}^{prey}}{f_i^{prey} \cdot f_j^{prey} \cdot \sum_{all-baits} n_{prey} \cdot \frac{(n_{prey}-1)}{2}}\right). \tag{A.8}$$

For the terms in $S_{i,j|i=bait}$ we have the following: $n_{i,j|i=bait}$ is the number of j's that i pulls down when i is bait, f_i^{bait} is the frequency that i was bait, n_{bait} is the number of unique bait proteins, f_j^{prey} is the fraction of times that j was pulled down by any bait that was not j itself, and $n_{i=bait}^{prey}$ is the total number of proteins i pulled down not counting itself. Since we are performing TAP *in silico*, we are guaranteed to retrieve all prey proteins for a particular bait and do not need multiple replications with the same bait (as all preys are explicitly accounted for in the snapshot). This means that $f_i^{bait} = \frac{1}{n_{bait}}$ and the S term simplifies to:

$$S_{i,j|i=bait} = \log\left(\frac{n_{i,j|i=bait}}{f_j^{prey} \cdot n_{i=bait}^{prey}}\right). \tag{A.9}$$

This term is thus the logarithm of the number of times i pulls down j, divided by the expected value of this number (which is the total number of times j is pulled down times the total number of proteins i pulls down as bait).

The terms in $M_{i,j}$ are as follows: $n_{i,j}^{prey}$ is the number of “purifications" in which i and j are observed together when neither i nor j are bait, f_i^{prey} and f_j^{prey} are the fraction of unique monomers pulled down by i or j, respectively, and n_{prey} is the number of unique monomers pulled down with bait i. Note that this last term is summed over all bait proteins, and thus does not require an index. $M_{i,j}$ is thus the logarithm of the observed “co-purifications" of i and j over its expected value,

which is the frequency of observing i and j together over all baits when neither i nor j are baits.

We created an $N \times N$ matrix of SA scores over all protein types ($N = 18$ unique proteins) for snapshots generated during peak signaling ($t = 360$ s) in both the machine and ensemble models. As this is a symmetric matrix, there are 153 unique pairings of proteins (and thus 153 SA scores). The majority of these pairings result in a score of 0 since there is no possibility of their presence in the same complex (*e.g.* Pheromone and Dig1). It is plain to see, however, that the SA scores which do exist are between proteins that are in the same bins as discussed above in Section A.3.4. The ensemble model's SA matrix is divided over two tables (Table A.14 and Table A.15); some interactions' SA scores are shown twice (*e.g.* Ste11 and Ste7) and the scores not shown are equal to 0. Fig. 2.6A in the main text shows the correlation between the values in the machine and ensemble SA matrices.

We can then create clusters or "functional modules" as referred to in [67] using the Markov clustering (MCL) algorithm outlined in [68]. The MCL algorithm partitions the set of proteins into disjoint clusters based on their SA scores, yet we know that certain proteins may associate with multiple types of complexes (*e.g.* Ste4 associates with G-protein related proteins and scaffold related proteins). To allow these proteins to be "shared" between modules we adapted Pu *et al.*'s method [67] and checked for proteins that had interactions with proteins in a distinct cluster. If a protein has positive SA scores with 75% of those in the "acceptor" cluster, we consider it a member of the acceptor cluster in addition to its original cluster. Representative clusters can be seen in Fig. 2.6A in the main text.

	Phe.	**Ste2**	**Sst2**	**Gpa1**	**Ste4**	**Ste20**	**Ste5**	**Ste7**	**Ste11**	**Ste12**	**Dig1**	**Dig2**	**Fus3**	**Kss1**
Phe.	0	5.1048	4.6352	4.2624	2.7848	0	0	0	0	0	0	0	1.8951	1.1902
Ste2	5.1048	0	4.6607	4.2888	2.8293	0	0	0	0	0	0	0	1.918	1.2127
Sst2	4.6352	4.6607	0	3.818	2.4891	0	0	0	0	0	0	0	2.9292	2.213
Gpa1	4.2624	4.2888	3.818	0	4.7597	0	0	0	0	0	0	0	1.0759	0.3699
Ste4	2.7848	2.8293	2.4891	4.7597	0	3.7073	3.5454	1.6443	2.8402	0	0	0	1.8825	1.3433
Ste20	0	0	0	0	3.7073	0	7.1188	5.1865	6.3936	0	0	0	3.6351	3.1986
Ste5	0	0	0	0	3.5454	7.1188	0	5.9115	7.0845	0	0	0	4.2624	3.8706
Ste7	0	0	0	0	1.6443	5.1865	5.9115	0	5.1393	0	0	0	4.4198	5.2097
Ste11	0	0	0	0	2.8402	6.3936	7.0845	5.1393	0	0	0	0	4.0232	3.6254
Ste12	0	0	0	0	0	0	0	0	0	0	9.2719	8.7543	3.939	5.0572
Dig1	0	0	0	0	0	0	0	0	0	9.2719	0	8.9105	3.8773	4.7526
Dig2	0	0	0	0	0	0	0	0	0	8.7543	8.9105	0	3.805	4.2255
Fus3	1.8951	1.918	2.9292	1.0759	1.8825	3.6351	4.2624	4.4198	4.0232	3.939	3.8773	3.805	0	1.5936
Kss1	1.1902	1.2127	2.213	0.3699	1.3433	3.1986	3.8706	5.2097	3.6254	5.0572	4.7526	4.2255	1.5936	0

Table A.14: Socio-affinity score table for proteins associated with the G-protein cycle, scaffold-based signaling, and transcriptional regulation. The three blocks of nonzero values correspond to the bins described in Section A.3.4 (note that the MAPKs are present all these bins).

	Ste7	Ste11	Mekp	Mekkp	Ptp	Msg5	Fus3	Kss1
Ste7	0	5.1393	9.711	0	0	0	4.4198	5.2097
Ste11	5.1393	0	0	9.7515	0	0	4.0232	3.6254
Mekp	9.711	0	0	0	0	0	3.5648	4.8434
Mekkp	0	9.7515	0	0	0	0	0.5554	0.1187
Ptp	0	0	0	0	0	0	2.5208	6.4925
Msg5	0	0	0	0	0	0	3.0833	6.4005
Fus3	4.4198	4.0232	3.5648	0.5554	2.5208	3.0833	0	1.5936
Kss1	5.2097	3.6254	4.8434	0.1187	6.4925	6.4005	1.5936	0

Table A.15: Socio-affinity score table (second)

A.3.7 Robustness of combinatorial inhibition

In order to confirm that our results on combinatorial inhibition were not artifacts of the parameter sets of the machine and ensemble models, we created 100 models with randomized parameters for both the machine and ensemble model. The procedure for generation and simulation of these models follows that described in Section A.2.4. We simulated these models at wild-type, 12x, and 60x concentrations of Ste5. We found that for all ensemble-based models, 12x concentrations of Ste5 increased Fus3 activation and 60x concentrations of Ste5 decreased Fus3 activation relative to the 12x activation level, confirming the robust presence of combinatorial inhibition (Fig. A.16). The machine-based models display resistance to combinatorial inhibition (Fig. A.17), with most models producing similar levels of Fus3 activation at 12x and 60x concentrations of Ste5.

Calculation of relative ΔFus3pp values in Fig. 2.6C of the main text was performed by subtracting the number of active Fus3 molecules at peak signaling (Fus3pp) for some scaffold concentration from Fus3pp for some higher scaffold concentration. In the case of Fig. 2.6C in the main text, these values are 60x WT and 12x WT. This value was then divided by Fus3pp for the lesser of the two scaffold values. The resulting value is then a measure of the relative increase (> 1) or decrease (< 1) of the change in Fus3 activation during peak signaling.

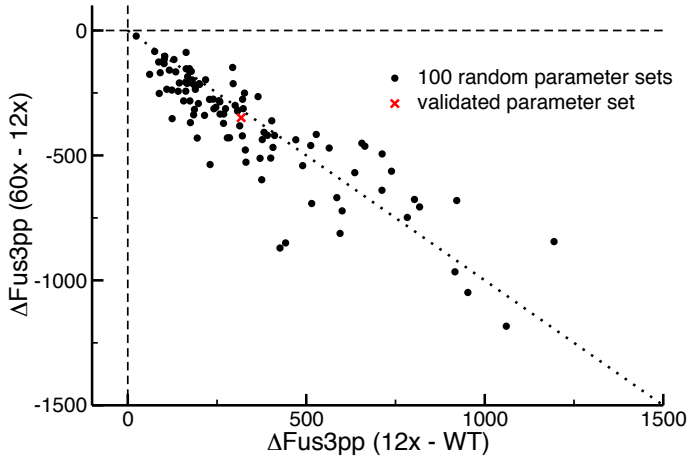


Figure A.16: Ensemble parameter randomizations (100 parameter sets). The x-axis is the difference in Fus3 activation (in number of molecules) between WT and 12x concentrations of scaffold and y-axis is the difference in Fus3 activation between 12x and 60x concentrations. Dashed lines are $x = 0$ and $y = 0$ for reference, and the dotted line is $y = -x$ to accentuate the strong correlation between these values.

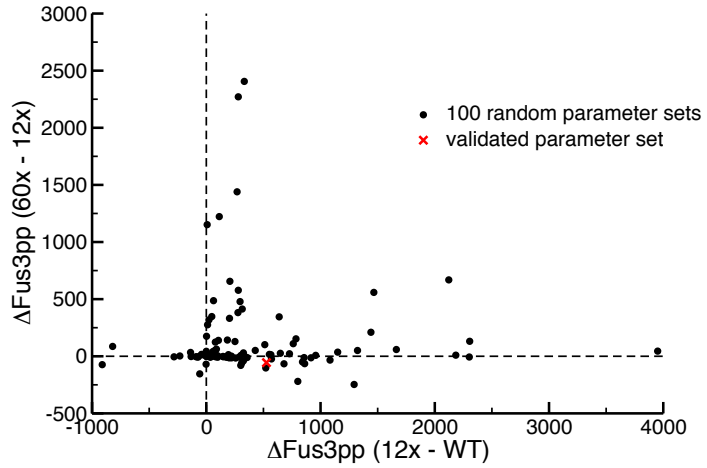


Figure A.17: Machine prozone parameter randomizations (100 parameter sets). Axes and dashed lines are as Fig. A.16

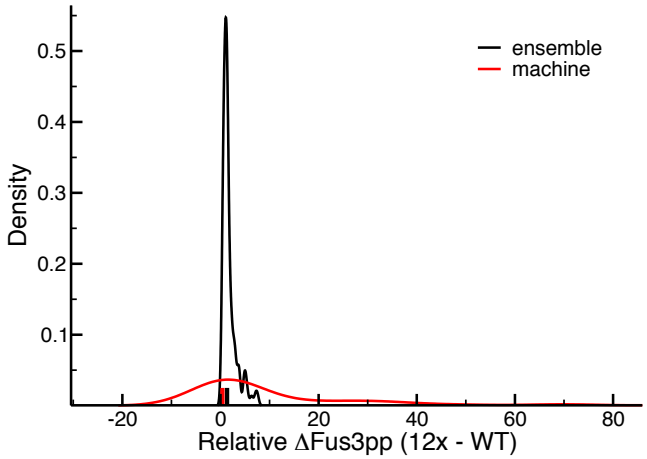


Figure A.18: Relative ΔFus3pp values for randomized machine and ensemble models have similar means when considering the difference in scaffold concentration between 12x and wild-type values, but the machine model's distribution has a much higher variance. The majority of values in both models, however, were positive, indicating a general increase in signaling activity.

Appendix B

Appendix for Chapter 2

B.1 Varying signal strength

Initially, we varied the copy number of the signaling molecule (K_0, representing an active signal molecule) to generate dose-response trends in our three models, however, a single such molecule induced a greater-than-negligible response. Therefore we implemented an alternate strategy for fully exploring the low-signal range of dose-response behavior of these systems, in which the activity of the first phosphatase in the cascade (P_1) was varied to suppress or allow signal throughput. Specifically, we modified the catalytic rate of P_1 (or $p_{cat,1}$) while simultaneously varying the association rate between P_1 and active K_1 $(p_{on,1})$ in order to maintain identical K_M values for all enzyme types:

$$p_{on,1} = \frac{p_{off} + p_{cat,1}}{K_M}. \tag{B.1}$$

This provides a fixed saturation level for this initial covalent modification cycle (here termed a Goldbeter-Koshland, or *GK*, loop [87]) that is identical to the maximum saturation levels in the subsequent substrate-modification reactions. It is also important to note that this first cycle is independent of any scaffold interaction, and is thus a true GK loop. Varying $p_{on,1}$ and $p_{cat,1}$ allows us to use the ratio of maximum enzyme velocities $\left(S = \frac{V_{max,K_0}}{V_{max,P_1}}\right)$ for this first GK loop to measure a simulation's exposure to signal. [86, 87].

We chose ranges of signal values empirically for each signaling paradigm such that both low- and high-response behaviors were observed, characterizing both minimum and maximum signal throughput. For the unsaturated ensemble and solution models and all saturated models, we varied S between $10^{-5} - 10^{5}$ s^{-1} in logarithmic increments. The unsaturated machine models exhibited stronger sensitivity to signal (measured as S_{50}), thus we examined a signal range of $10^{-8} - 10^{2}$ s^{-1} with identical increments.

B.2 Unsaturated models

In the main text, our results focused exclusively on models with a 1:10 phosphatase to kinase copy number ratio (with the exception of the first and last kinases as mentioned in section B.1 and in the main text, respectively). As expected, varying this quantity influences the qualitative trends minimally and merely inhibits signal throughput at greater cascade depths as seen in Figure B.1.

B.2.1 Signal amplification

Since scaffold proteins have been proposed to decrease signal amplification [22, 23, 80], we proceeded to investigate these claims within our three signaling paradigms. We defined amplification as the ratio of initial kinase activation to final kinase activation: $\frac{K_1^*}{K_F^*}$. Since the activation of the first kinase in all signaling paradigms is equivalent to the activation of the substrate in an isolated GK loop (see section B.1), we can analytically calculate this quantity and apply it to our calculation of the signal amplification. Since the kinases in our models are not saturated (including the initial kinase) we cannot employ the typical Michaelis-Menten approximations and instead use the total quasi-steady-state approximation outlined in [136] to determine the level of active initial kinase. Depth-dependent amplification for each signaling paradigm (with a phosphatase to kinase ratio of 1:10) can been seen in Figures B.2-B.4.

A number of trends emerge, most notably that all paradigms exhibit at least 30x peak amplification for some level of signal. Furthermore, optimal amplification takes place when the signal levels

are moderately low and increased cascade depth generally corresponds to increased amplification. We can therefore conclude that, in our models, scaffolding does not preclude signal amplification, and the nature of the scaffold-mediated signaling (*i.e.* machine- or ensemble-like) greatly impacts the level of amplification.

B.2.2 Varying scaffold number

Figures B.5-B.7 show Hill parameter trends and CoV trends mentioned in the main text.

B.3 Saturated models

Shown clearly in Figure B.8, we observe minimal difference between the three signaling paradigms with saturating conditions. There are slight qualitative differences between varying phosphatase levels, such as the variation in depth-dependent signal sensitivity trends (*i.e.* low signal throughput results in slightly less sensitive response for mid to large cascade depths, and even less sensitivity for shallow cascades; Figure B.9). However the Hill-based parameters generally reflect similar behavior across all paradigms, which stems from our assumption of noncompetition between substrates and phosphatases for kinases. As a result, the initial saturated modification cycle (*i.e.* K_0 binding and activating K_1) follows precisely the dynamics of a typical Goldbeter-Koshland futile cycle [87]. The resulting ultrasensitive behavior then propagates through the subsequent covalent modification cycles (themselves similarly ultrasensitive) producing a switch-like response from the final kinase (Figure B.10). Since this behavior is consistent throughout all three examined signaling paradigms, we chose to focus on the unsaturated parameter regime due to its relative phenotypic diversity.

As a side note, over the course of our calculations, we observe jagged edges in our matplotlib-based interpolation of the saturated data sets. This is due to the relatively coarse-grained sampling of both phosphatase concentration space as well as signal space; this prevented accurate fitting of some data sets to the Hill function (see main text) and still more were thrown out due to insignifi-

cant parameterization.. Since the saturated models exhibit such sharp ultrasensitivity (Figure B.10), it would take extensive sampling of parameter space to produce smoother graphs (and is outside the scope of this work).

B.4 Combinatorial complexity in species formation

As mentioned in the main text, the ensemble model exhibits far greater combinatorial complexity in terms of signaling species formation. This was initially examined in [1, 11]. We observe this phenomenon in Figure B.11 where, at increased cascade depths, ensemble-like networks are capable of sampling over an order of magnitude more species than their machine-like counterparts. The solution model, due to its lack of scaffold, reveals even less complexity in its network; at the greatest examined cascade depth, ensemble-like networks can produce over three orders of magnitude more signaling species. This result is quite striking, especially when considering that the machine and ensemble signaling paradigms generate far less intrinsic noise.

B.5 Causality analysis

In order to determine how the ensemble models reduced inappropriate pathway output in our analysis of crosstalk, we proceeded to examine the trajectory of events necessary for activation of the outputs of pathway A and B when only pathway A's signal was active. We therefore produced strongly compressed causal flows, termed *causal histories* or *stories*, which are formal mathematical representations of how a particular structure is constructed [92]. Upon generation of stories for $K_{3,A}$ and $K_{3,B}$ when pathway A is maximally active and pathway B is minimally active (Figure B.12) we observe that the story for $K_{3,B}$ (right) involves 2 more events than that for $K_{3,A}$ (left). Specifically, $K_{3,B}$ activation in this system requires that active K_2 unbinds $Scaf_A$ and subsequently binds $Scaf_B$. These additional events are responsible for the reduction in pathway B activation relative to the solution model's equal activation of both $K_{3,A}$ and $K_{3,B}$ when only pathway A is active.

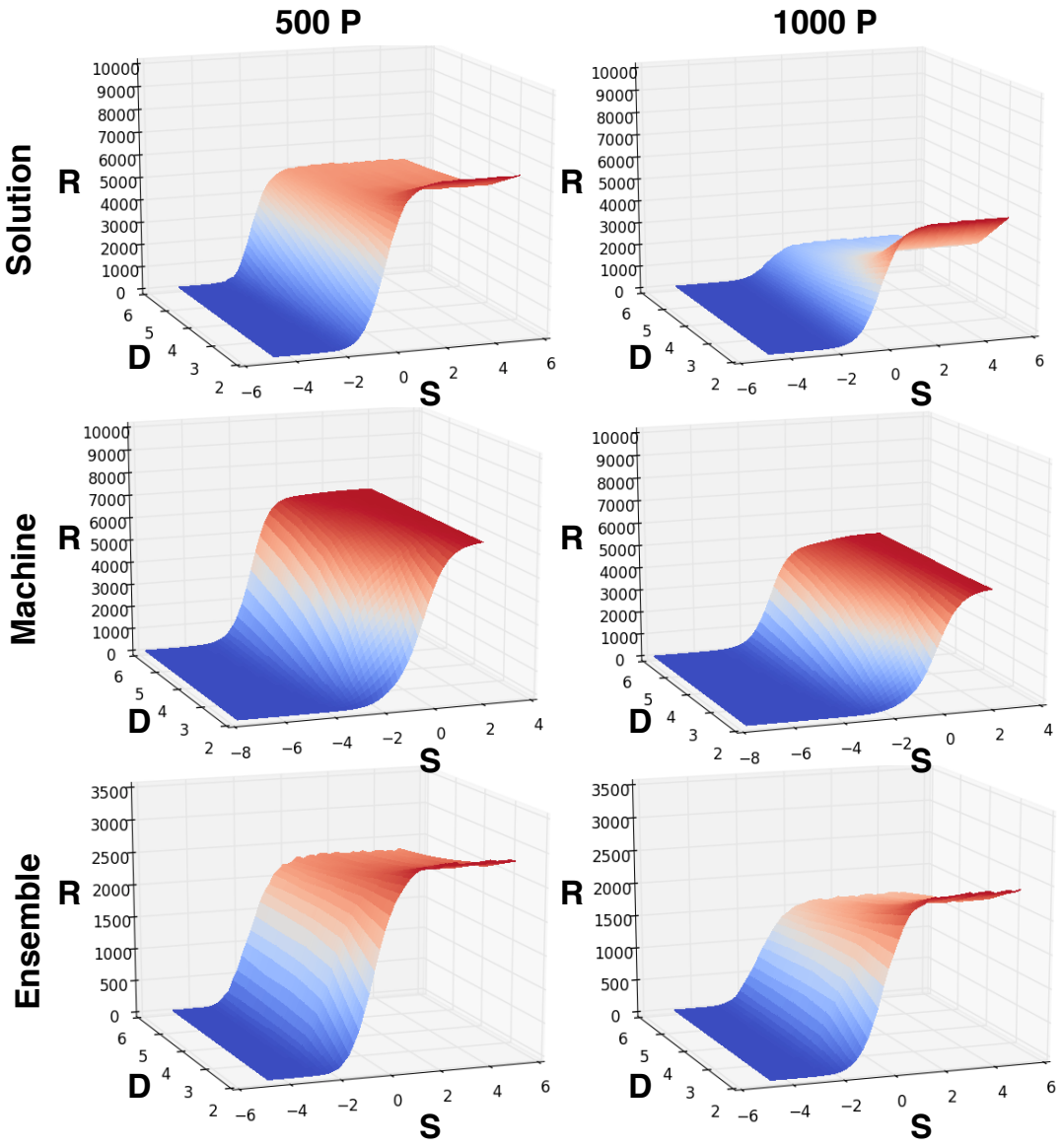


Figure B.1: Dose response trends for select unsaturated models. The left column of models contains those with a 2:1 phosphatase to kinase ratio, and the right column has a ratio of 1:1. We observe a decrease in response (**R**) for cascades of greater depth (**D**), though the qualitative signal (**S**)-dependent trends remain consistent across parameter space.

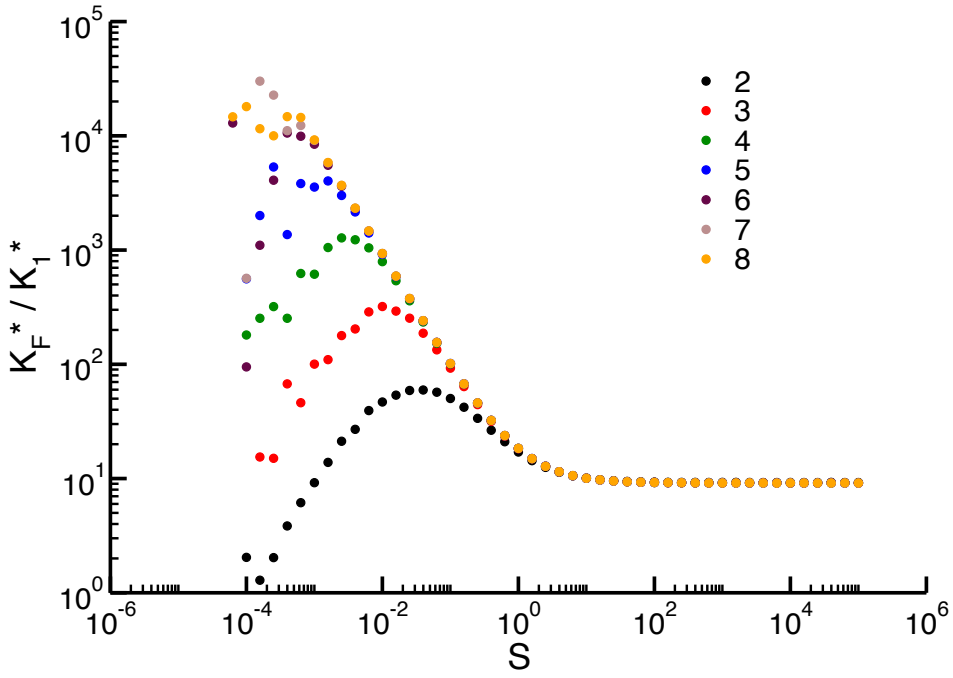


Figure B.2: Signal amplification dependence on signal level in solution models of varying depth.

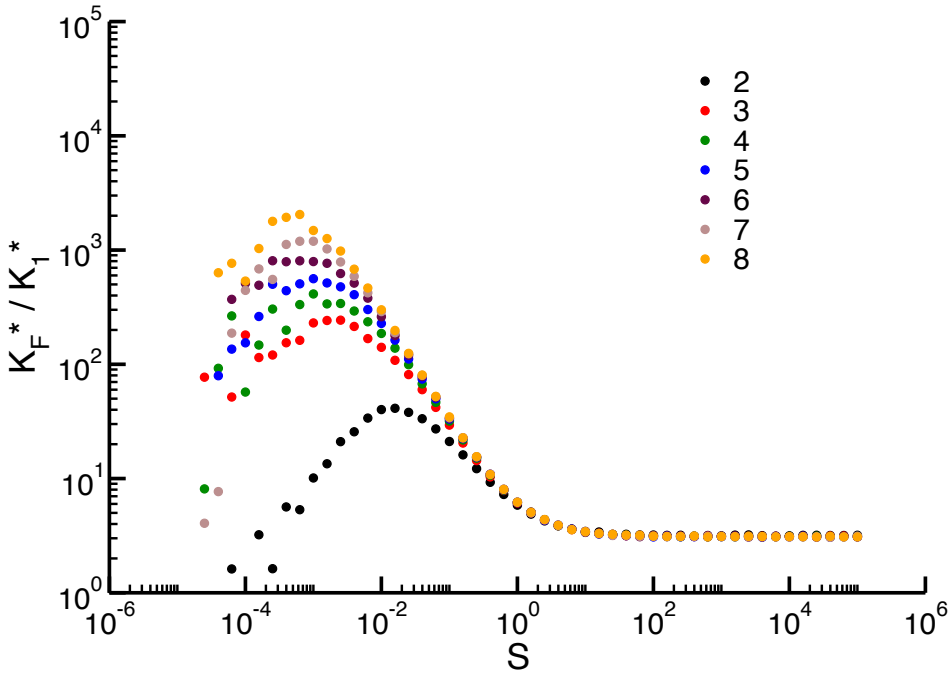


Figure B.3: Signal amplification in ensemble models.

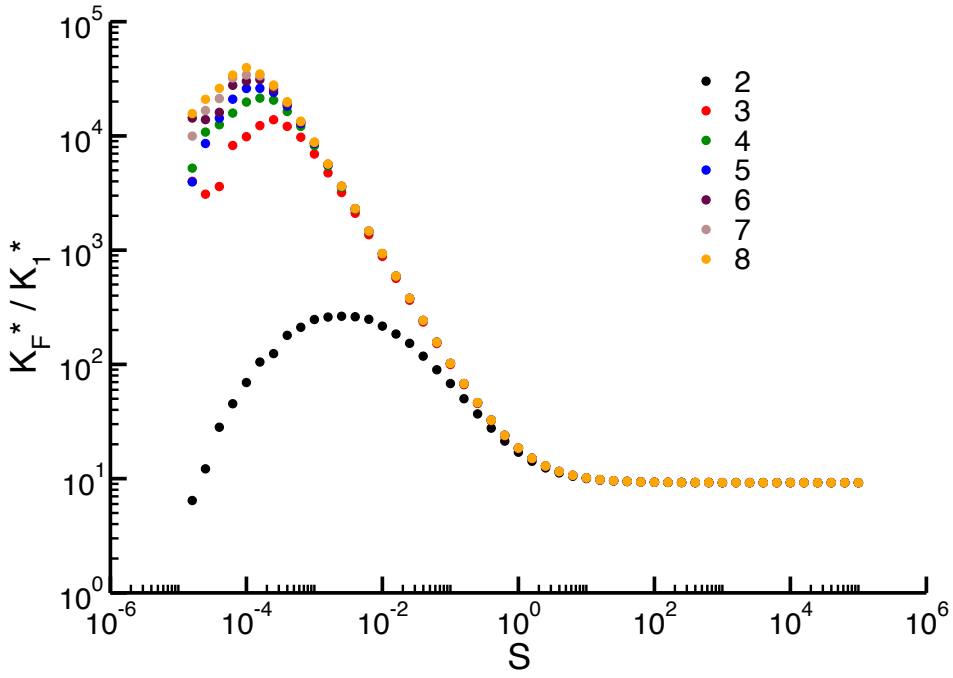


Figure B.4: Signal amplification in machine models.

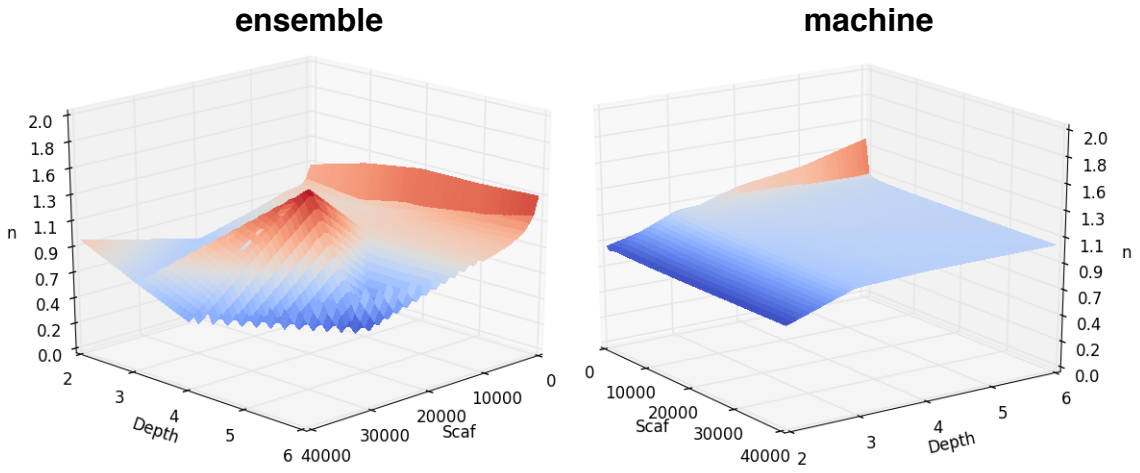


Figure B.5: Dose-response ultrasensitivity as a function of depth and scaffold number. For both machine and ensemble models, a decrease in ultrasensitivity is observed with increasing scaffold number.

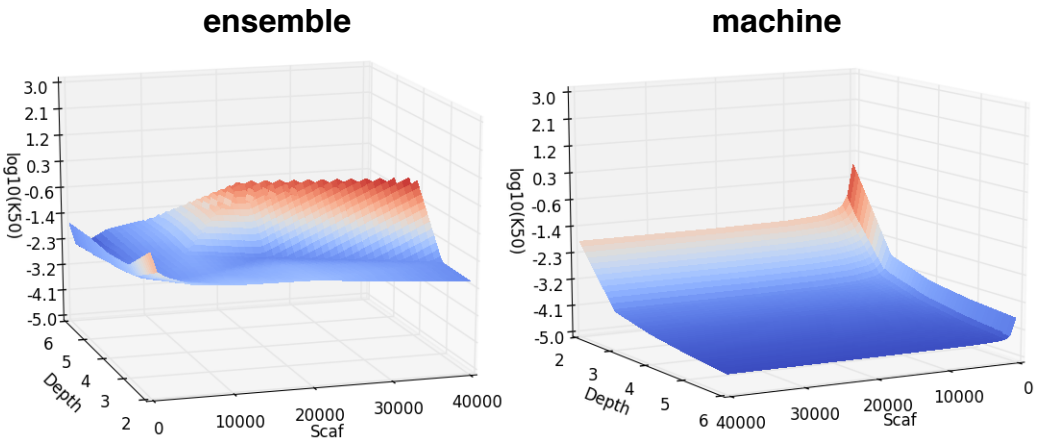


Figure B.6: Sensitivity to signal with respect to depth and scaffold number. Sensitivity increases to a certain point and remains constant in the machine model. This also occurs in the ensemble model, however once the effects of combinatorial inhibition become relevant (at higher scaffold numbers) the sensitivity to signal decreases.

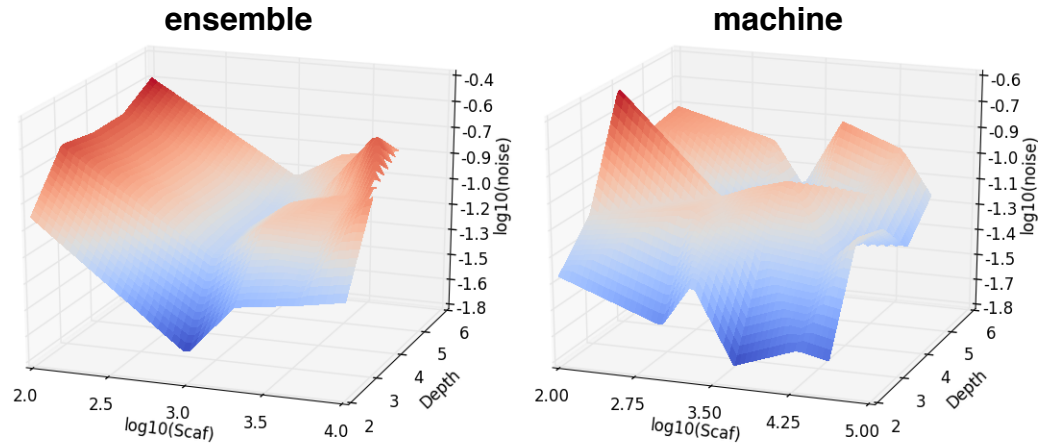


Figure B.7: The ensemble model exhibits a minimum amount of noise when the scaffold and kinases are at stoichiometric ratios, though the increase resulting from deviating from this value is less than an order of magnitude in the sample parameter space. The machine model does not exhibit such a clear trend, however the fluctuations in noise are again less than an order of magnitude from the default parameter set.

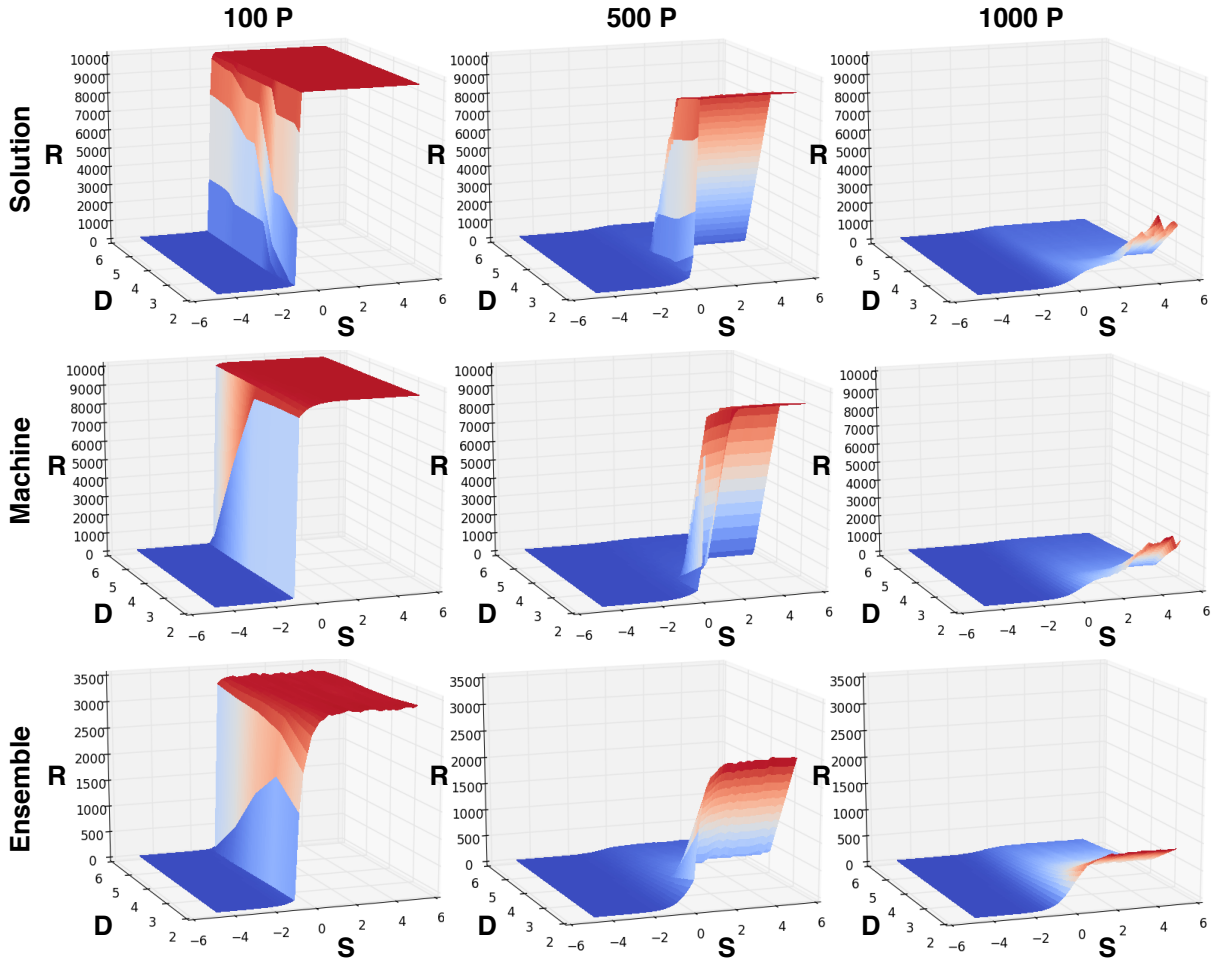


Figure B.8: Dose response trends for select saturated models. From left to right, the phosphatase to kinase ratio is 1:10, 1:2, 1:1 and the axis labels are identical to those in Figure B.1. We observe similar behavior to that seen in unsaturated models despite the fact that the transitions between the inactive and active regimes of the cascades occur much more sharply.

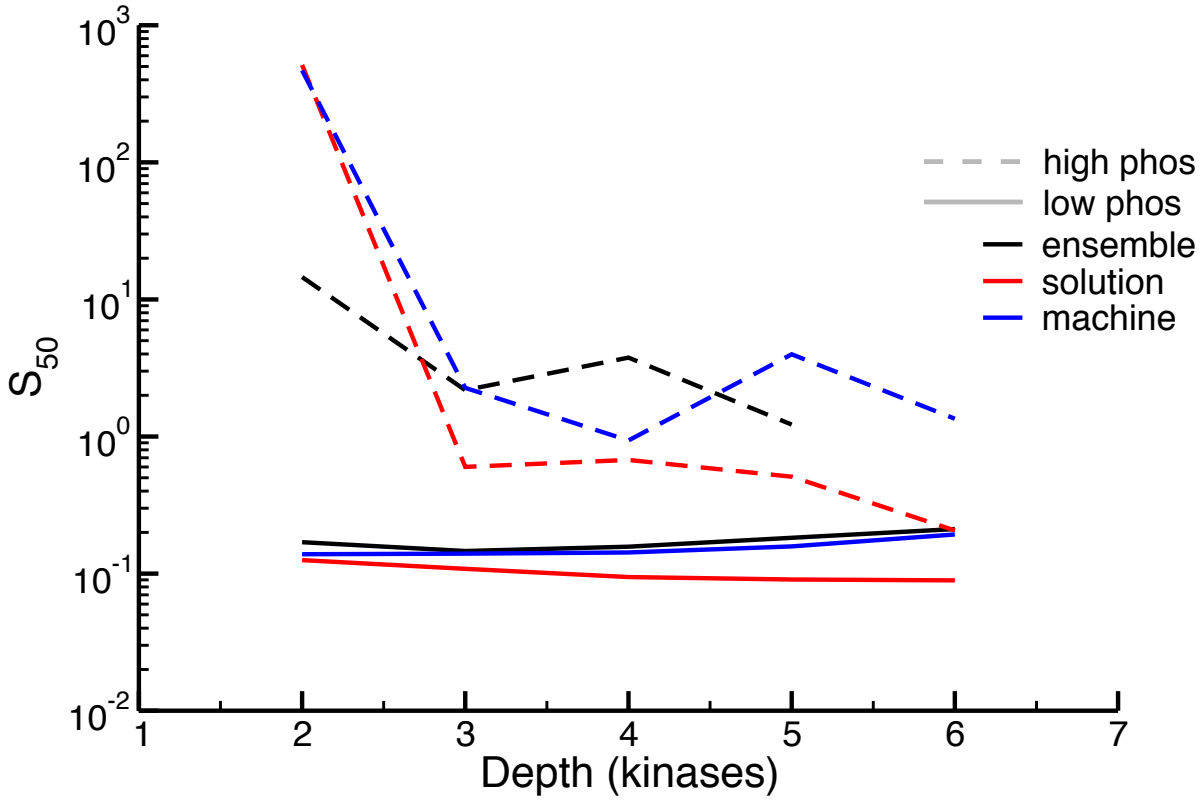


Figure B.9: The saturated models exhibit notably lower sensitivity to signal. The most major qualitative distinction between these and the unsaturated models' trends is the invariant behavior of the models with respect to cascade depth (with the exception of the 2-kinase cascades).

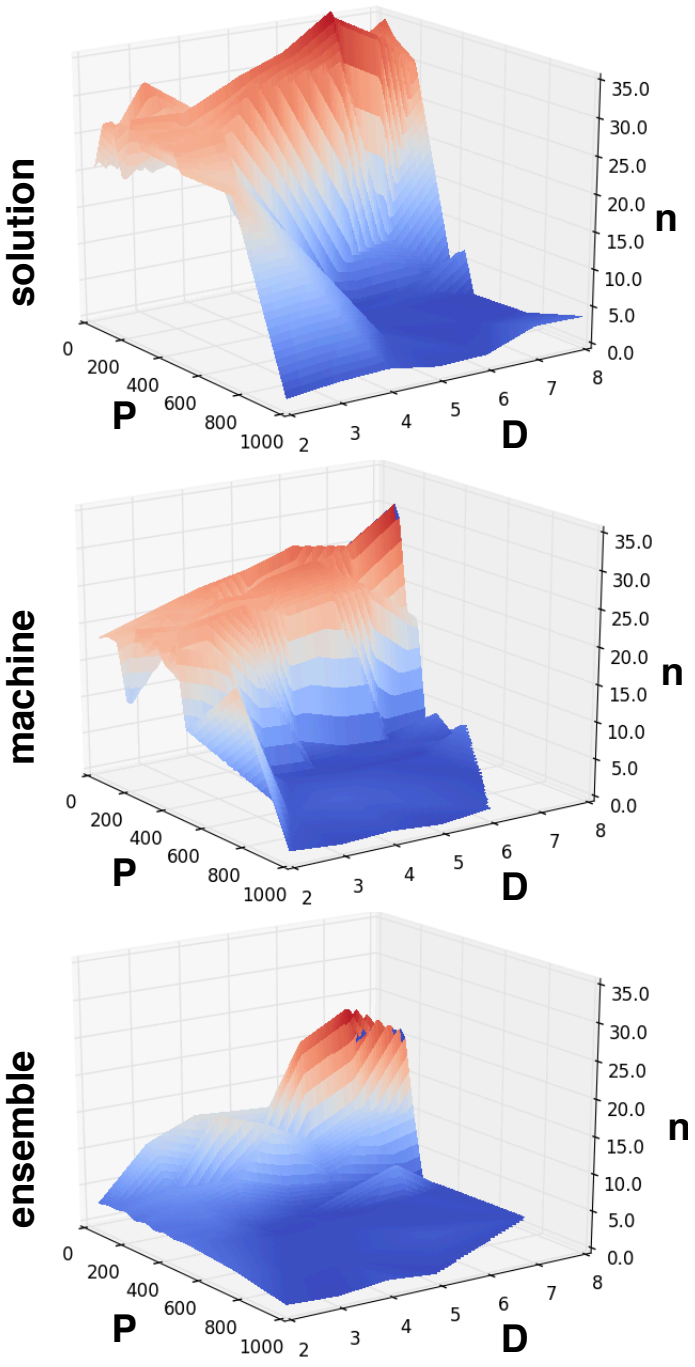


Figure B.10: The saturated models are much more ultrasensitive relative to the unsaturated models. Here, **P** is the phosphatase concentration, **n** (or n) is the Hill coefficient, and **D** is cascade depth. The machine and solution models have similar Hill coefficients at higher depths, and although the ensemble model maintains lower values across the searched parameter space, we observe notably larger n values as compared to the relevant unsaturated models.

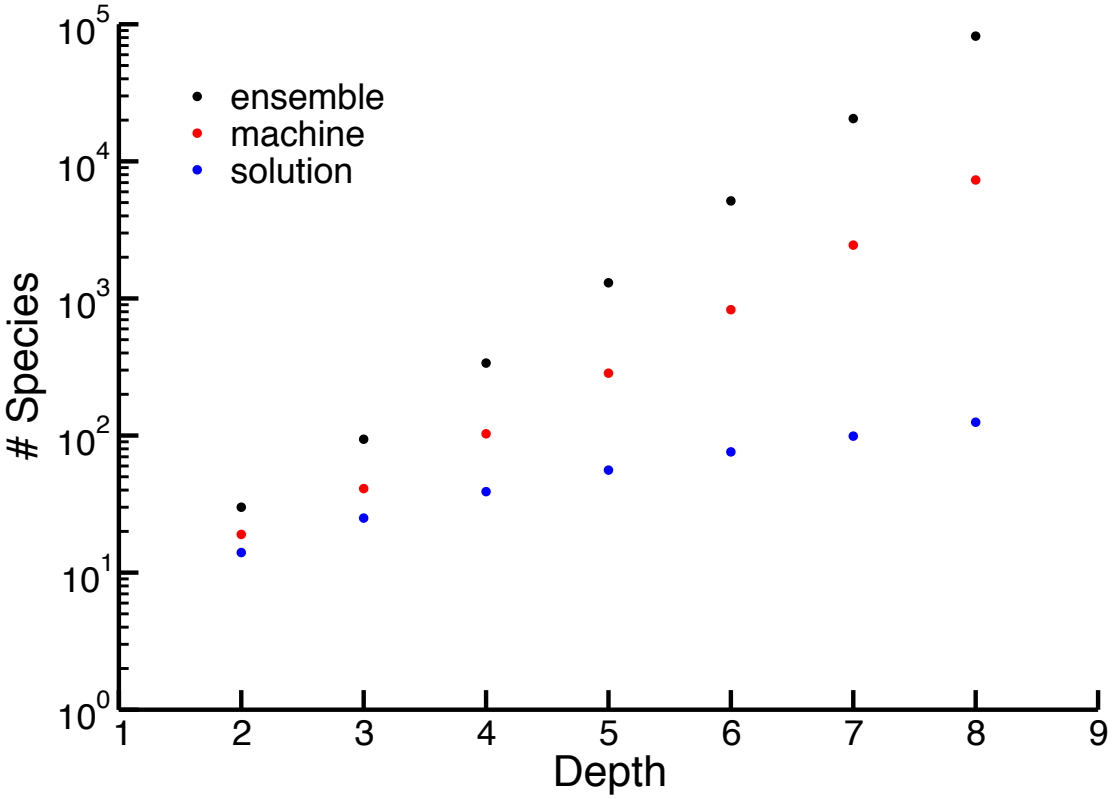


Figure B.11: Combinatorial complexity in the ensemble model (black) leads to an increased number of potential species as compared to the machine (red) and solution (blue) models. This phenomenon is exaggerated as the depth of the cascade is increased.

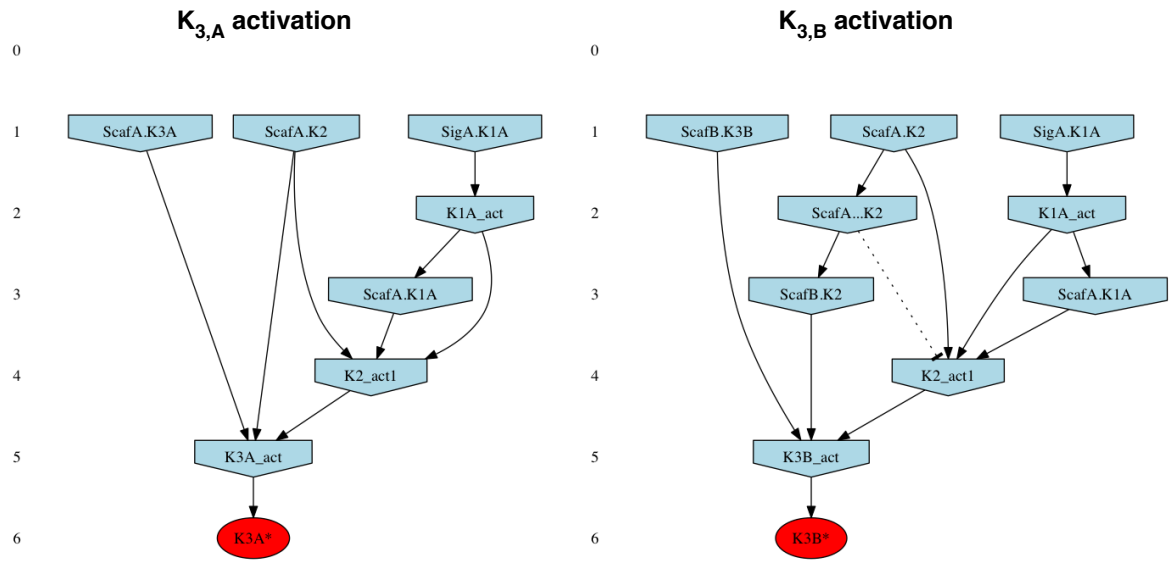


Figure B.12: Output of strongly compressed causal histories from KaSim, showing the causal relationships between the rules relevant for $K_{3,A}$ and $K_{3,B}$ activation when only pathway A is active. Each blue node in the directed story graph is an event that is required for formation of the final object of interest, in this case, activation of the kinase following the shared kinase (K_2) in either pathway A or B (red nodes). The edges in this graph represent either an activating (solid with arrows) or inhibiting (dotted with bars) influence between the upstream and downstream events.

Appendix C

Appendix for Chapter 3

C.1 Information Theory Calculations

C.1.1 Mutual Information

The mutual information between two random variables representing a signal, S, and a response, R, is defined as:

$$I(S;R) = \int_S \int_R p(s,r) \log \frac{p(s,r)}{p(s)p(r)} \mathrm{d}s\mathrm{d}r, \tag{C.1}$$

where S is a random variable representing the input signal, R a random variable representing the response, $p(s,r)$ is the joint probability distribution for some combination of s and r values, and $p(s)$ and $p(r)$ are the corresponding marginal distributions [26]. One of the major difficulties in calculating this quantity from experimental data is the fact that the continuous probability density functions defined above must be estimated on the basis of an inherently discrete data set. As a result, a number of approaches have been developed to obtain unbiased estimates of the mutual information with varying degrees of accuracy [117].

In order to facilitate comparison with earlier results, we employed the same strategy used by Cheong *et al.* [9]. This strategy has two main components. First, one defines a set number of “bins" in both the signal values s and response values r. In cases where one is measuring the molecular

response of individual cells to a given signal (e.g. nuclear localization of NF-κB upon treatment with TNF-α, [9]), there are a small number of ligand concentrations used to treat the cells, resulting in a natural discretization of the S variable and a total of S_B bins of signal values. One defines a number of bins for the response (R_B), and uses these bins to estimate the probability of observing some response bin given some signal bin (i.e. $p(r|s)$). A linear extrapolation procedure is then used to estimate the mutual information one would obtain if there were an infinite amount of data in the data set. This extrapolation procedure is described in greater detail in section C.1.1.2 below.

One issue with this approach, however, is that the number of bins into which the signal and response values should be divided is not well-defined; using a larger number of bins generally increases the estimated amount of information [9]. To combat the potential for overestimation of the mutual information, the second phase of the procedure involves varying the total number of bins in the response variable (and, when appropriate, in the signal value as well) and estimating I for both the experimental data and a set of randomized replicates of the data. This allows one to choose a bin size that maximizes I for the real data while still estimating 0 information for the randomized versions. This element of the procedure is detailed in section C.1.1.3. Estimates of the mutual information based on this approach can subsequently be used to calculate the channel capacity by finding the input distribution that maximizes I [26] (section C.1.2).

C.1.1.1 Calculating the mutual information

To calculate the mutual information from our finite data sets, we first created a “contingency table" K based on the data: the rows of this matrix represent the various signal bins, and the columns are the various response bins. Each entry in the matrix is the number of observations from the data that correspond to that particular signal-response pair. The contingency table for a particular experiment might look something like this:

$$K = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{array} \begin{array}{c} \begin{array}{ccccc} r_1 & r_2 & r_3 & r_4 & r_5 \end{array} \\ \begin{pmatrix} 6 & 1 & 0 & 0 & 0 \\ 0 & 3 & 4 & 0 & 0 \\ 0 & 1 & 2 & 4 & 0 \\ 0 & 0 & 0 & 2 & 5 \end{pmatrix} \end{array}$$

Note that the above table is meant only as an example, and does not contain actual data. One can use the contingency table to calculate the mutual information in terms of the marginal and conditional entropies:

$$I(S;R) = H(S) - H(S|R) \tag{C.2}$$

$$I(S;R) = -\sum_i^{S_B} p(s_i) \log p(s_i) - \sum_i^{S_B} \sum_j^{R_B} p(s_i, r_j) \log \frac{p(r_j)}{p(s_i, r_j)} \tag{C.3}$$

where i ranges over the signal bins and j over the response bins in the contingency table (recall that S_B and R_B are the total number of signal and response bins, respectively). Since each entry in the contingency table can be naturally considered a conditional probability, it is helpful to rewrite this equation as:

$$I(S;R) = -\sum_i^{S_B} p(s_i) \log p(s_i) + \sum_j^{R_B} p(r_j) \sum_i^{S_B} p(r_j|s_i) \log p(r_j|s_i). \tag{C.4}$$

We can then calculate the frequencies from the contingency table entries and substitute these values into the equation. We define N_T as the sum over all entries in the table. Since each entry of the matrix, k_{ij}, is the number of instances of signal i that resulted in response j, we can define the total number of observations corresponding to a given signal bin i as $k_{s,i} \equiv \sum_j^{R_B} k_{ij}$. Similarly, we can define the total number of times any particular response bin j was observed as $k_{r,j} \equiv \sum_i^{S_B} k_{ij}$. Given these definitions, we can calculate the mutual information using the following equation:

$$I(S;R) = -\sum_i^{S_B} \frac{k_{s,i}}{N_T} \log \frac{k_{s,i}}{N_T} + \sum_j^{R_B} \frac{k_{r,j}}{N_T} \sum_i^{S_B} \frac{k_{ij}}{N_T} \log \frac{k_{ij}}{N_T}. \tag{C.5}$$

Equation C.5 is used whenever a particular value of I is calculated in the estimation procedure described below [9, 117].

C.1.1.2 Removing bias due to finite sample size

Although it is straightforward to use equation C.5 to calculate the mutual information, the fact that there are a finite number of data points in the contingency table (N_T) can introduce biases into the calculation. To estimate this bias, one can create a smaller data set with N_T' points ($N_T' < N_T$) and calculate I. As has been observed previously [9], as N_T' decreases, bootstrap replicates of the data generate higher values of I. This results in a roughly linear decrease in I as the inverse sample size decreases (Figure C.1).

To correct for this bias, we used the linear extrapolation procedure employed in Cheong *et al.* and other previous studies [9, 119]. We chose a total of 5 sample sizes starting from 60% of the original data with uniform increments in inverse sample space until reaching the size of the original data set. Our procedure then involves calculating the joint frequency distribution for the original data set and then randomly sampling the specified number of values over 20 independent replicates. We used these randomly sampled data sets to generate new contingency tables, and using equation C.5 we calculated the distribution of I across the 20 replicates. We then performed a linear regression of the I *vs.* $1/N_T$ relationship (*e.g.* the straight lines in Figure C.1). The y-intercept of these lines represents the extrapolation to an infinite data set (i.e. $N_T \to \infty$ implies $1/N_T \to 0$). All the channel capacities calculated in this work (*e.g.* those reported in Table 1 of the main text) were obtained from these y-intercepts. The errors reported for these values in Table 1, and the error bars in all figures, represent 95% confidence intervals on the linear model's intercept estimate.

C.1.1.3 Finding the optimal number of bins

Generating the contingency table relies on a particular discretization or binning of the data. As mentioned above, the signal values used to generate experimental data often represent a natural

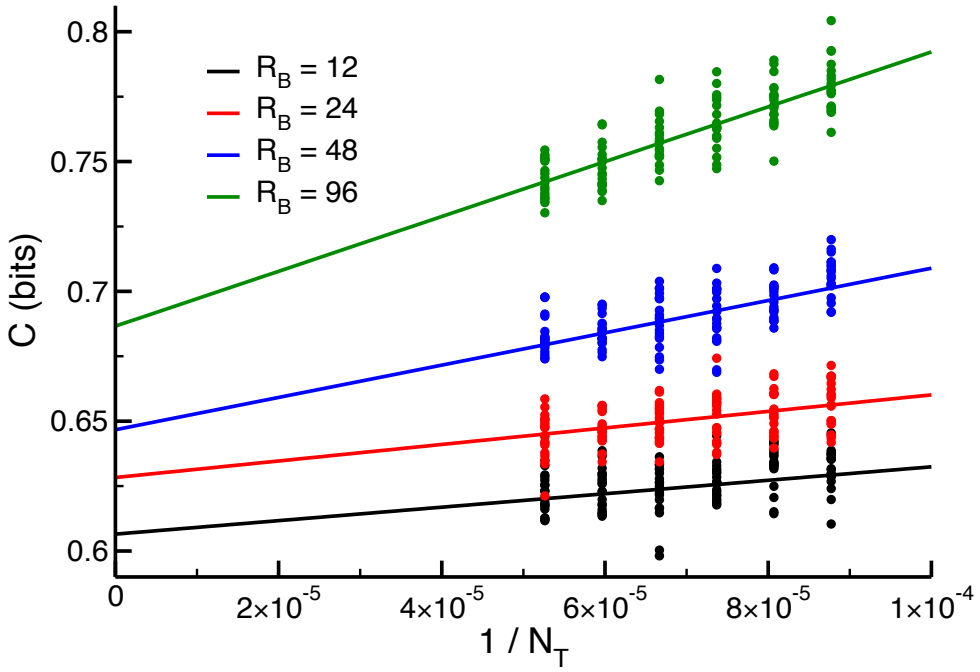


Figure C.1: Representative linear models for estimating mutual information at infinite sample size with various numbers of response bins. Here we use experimental data composed of 1000 cells per each of 19 TRAIL concentrations. In this figure we are estimating the mutual information between the level of caspase-8 activity in HeLa cells in response to a uniformly distributed set of TRAIL concentrations. We then calculate the mean mutual information as a function of inverse sample size ($\frac{1}{N_T}$) by taking $n = 20$ independent subsets of the data per sample size. Shown here are the calculated mutual information values for each sampled data set. Calculation of the linear model's intercept provides us with an estimate of mutual information at infinite sample size for a particular number of response bins.

set of signal bins (e.g. Figure 4.1C and D of the main text). The number of response bins to generate, however, is not clear *a priori*, and the value of R_B has a large impact on estimates of I. On one extreme, if we set $R_B = 1$, all of the signals will give the same responses, resulting in a mutual information of 0. Alternatively, we could choose a number of bins, R_B, so large that every response bin contains exactly one response value in the contingency table:

$$K' = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & r_7 & r_8 & r_9 & r_{10} & r_{11} & r_{12} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

(where again we have used an arbitrary data set as an example). This results in a (spuriously) high mutual information–note that, in this case, if we randomly shuffle the signal value that gives any particular output, we will get the same mutual information:

$$K'_{rand} = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & r_7 & r_8 & r_9 & r_{10} & r_{11} & r_{12} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Since I generally increases with an increasing R_B (note the increasing intercept for the data in Figure C.1), we must find an optimal value of R_B that accurately represents the mutual information in the underlying data without artificially inflating it.

Our approach to solving this problem is broadly inspired by previous approaches, particularly that of Cheong *et al.*, with some slight modifications [9, 116]. Each calculation of the mutual information requires some specified number of bins, so we defined both an initial number of bins to use as well as the number of bins by which to increment. For any given R_B value, we generated the bins themselves (*i.e.* the actual range of response values in the data that belongs to each bin)

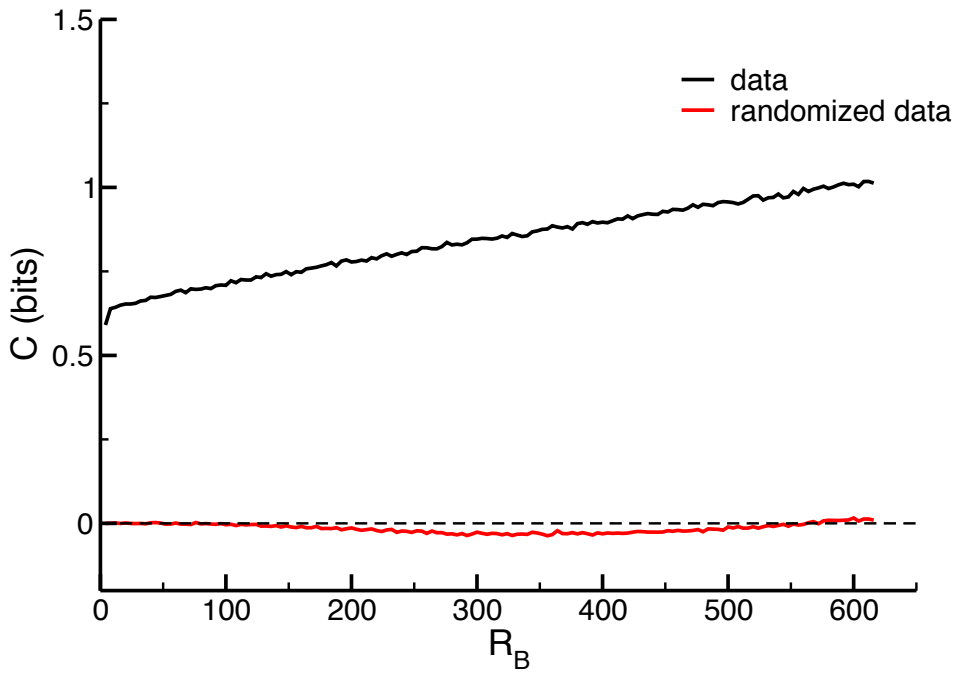


Figure C.2: Here we show a representative graph of mutual information (I) as a function of the number of response bins. The data is the same from which Figure C.1 is generated and each point shown here is the y-intercept retrieved from the linear extrapolation procedure outlined in section C.1.1.2. In red is one randomization of the actual data set, shown in black.

so that the total number of observations $k_{r,j}$ for each response bin is (roughly) equal across all the response bins under the signal distribution given by the data set [9, 116]. We then generated the contingency table and estimated I using the linear extrapolation procedure explained above in section C.1.1.2.

Plotting I *vs.* the total number of response bins (Figure C.2) does indeed demonstrate that mutual information increases essentially monotonically with increasing R_B. For each value of R_B, we also generated N_R contingency tables with some specified sample size with randomly sampled entries. We calculated I for each one of them; the value of I in these randomized data sets also increases with increasing R_B, eventually generating significantly non-zero mutual information where there should be none (Figure C.2).

Cheong *et al.* obtained an optimal range of bin numbers for each data set via visual inspection of plots like those in Figure C.2 [9]. While this is an effective approach, the large number of data sets and variants in our case prevented us from visually analyzing every case. We thus defined a uniform criterion for choosing the optimal number of bins, defined as the value of R_B that gives the largest value of I, subject to the constraint that the 95% confidence interval from the corresponding randomized data *must include* 0. In other words, we chose an R_B that maximizes the I in the data, but where the randomized data gives mutual information that is not significantly greater than 0.

The range of R_B values that provides this maximum depends on the total number of data points (N_T) and on the amount of information present in the data itself; it is thus difficult to define a uniform range of bin numbers to consider for every data set. Therefore we implemented a method to automate the search for the optimal number of bins. In essence, this method iteratively increments the number of bins used in the calculation until a prespecified number of randomized calculations (in our case, 3) for consecutive increments were biased (*i.e.* significantly above zero).

The discussion above assumes that S_B is fixed at some particular number of signal values, as is typical when generating experimental data. In some of the systems we considered, however, we needed to find an optimal set of signal bins in addition to response bins. This was particularly true of spatial quantities like the angle between the bacterium and the neutrophil (see section C.3). In those cases, we adapted this method to identify the optimal number of bins in both signal and response space.

C.1.2 Channel Capacity

As mentioned in the main text, the channel capacity is the supremum of the mutual information over all possible signal distributions. Estimating the channel capacity thus involves using the estimate of mutual information obtained from the procedure defined in section C.1.1 to search the space of signal distributions and find the one that maximizes I. Since the set of such distributions is obviously infinite, an exhaustive search of all well-defined signal distributions is impossible. Following the example of Cheong *et al.* [9], we implemented a grid-based search, limited to a

set of unimodal and bimodal Gaussian distributions in addition to a uniform distribution of signal values and the distribution present in the original data set if it is not uniform.

C.1.2.1 Unimodal signal distributions

We generated a range of unimodal signal distributions of the form:

$$G_U(s) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(s-\mu)^2}{2\sigma^2}}. \tag{C.6}$$

Since we can sample only a subset of possible signal distributions we limit the potential mean values, μ, to a set of 4 evenly spaced values between the minimum and maximum signal values in the data set (S_{min} and S_{max}, respectively):

$$\mu \in \{f \cdot (S_{max} - S_{min}) + S_{min}\} \mid f \in \{0.2, 0.4, 0.6, 0.8\}\}. \tag{C.7}$$

For each μ we generate a range of σ values by calculating a maximum standard deviation:

$$\sigma_{max} = \min\left(\frac{\mu - S_{min}}{3}, \frac{S_{max} - \mu}{3}\right). \tag{C.8}$$

This constrains the signal distributions so that at least 99% of the area under the distribution falls between S_{min} and S_{max} and allows us to use a range of standard deviation values by sampling increasing fractions of σ_{max}:

$$\sigma \in \{f \cdot \sigma_{max} \mid f \in \{0.2, 0.4, 0.6, 0.8, 1\}\}. \tag{C.9}$$

C.1.2.2 Bimodal signal distributions

We also implemented a range of bimodal signal distributions of the form:

$$G_B(s) = \frac{w_0}{\sigma_0\sqrt{2\pi}} e^{-\frac{(s-\mu_0)^2}{2\sigma_0^2}} + \frac{w_1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(s-\mu_1)^2}{2\sigma_1^2}} \tag{C.10}$$

where w_0 and w_1 are weighting coefficients such that $w_0 \in \{0.4, 0.5, 0.6\}$ and $w_1 = 1 - w_0$. In order to construct these distributions we first defined a minimum difference between μ_0 and μ_1:

$$\mu_D = \frac{S_{max} - S_{min}}{5}. \tag{C.11}$$

We used μ_D to construct a series of pairs (μ_0, μ_1) such that

$$\mu_D + S_{min} \leq \mu_0 < S_{max} \tag{C.12}$$

$$\mu_0 + \mu_D \leq \mu_1 < S_{max} \tag{C.13}$$

and μ_0 is incremented in steps of μ_D. Similarly to the unimodal signal distributions, both means μ_0 and μ_1 have multiple, evenly spaced standard deviations, σ_0 and σ_1 that are fractions of some maximum standard deviations, $\sigma_{0,max}$ and $\sigma_{1,max}$. These values are constrained so that these distributions have both a local minimum between μ_0 and μ_1 and 99% of their area between S_{min} and S_{max}:

$$\sigma_{0,max} = \min\left(\frac{\mu_1 - \mu_0}{4}, \frac{\mu_0 - S_{min}}{3}\right) \tag{C.14}$$

$$\sigma_{1,max} = \min\left(\frac{\mu_1 - \mu_0}{4}, \frac{S_{max} - \mu_1}{3}\right). \tag{C.15}$$

The individual σ_0 and σ_1 values are then:

$$\sigma_0 = f \cdot \sigma_{0,max} \tag{C.16}$$

$$\sigma_1 = f \cdot \sigma_{1,max}, \tag{C.17}$$

where $f \in \{0.2, 0.4, 0.6, 0.8, 1\}$.

C.1.2.3 Weighting the data

With this set of unimodal and bimodal signal distributions, we can determine how the mutual information of a particular data set varies with different signal distributions in order to estimate the channel capacity. To do this, we modified the original contingency table in order to recalculate the mutual information according to each new signal distribution. Each signal bin s_i corresponds to a range of signal values between, say, $s_{i,min}$ and $s_{i,max}$ and yields a corresponding number of observations in the contingency table, k_{ij} for each response bin r_j. For any unimodal or bimodal signal distribution $G_A(s)$, we calculated the new value for this entry in the contingency table $k'_{ij} = \frac{p'(s_i)k_{ij}}{p(s_i)}$ where

$$p'(s_i) = \int_{s_{i,min}}^{s_{i,max}} G_A(s)\,\mathrm{d}s \tag{C.18}$$

is the new probability of observing some signal value s_i and $p(s_i)$ is the uniform probability of observing that signal bin s_i. We can use this to generate a new contingency table, and calculate the relevant quantities:

$$N'_T = \sum_i^{S_B}\sum_j^{R_B} k'_{ij},\ k'_{s,i} = \sum_j^{R_B} k'_{ij},\ k'_{r,j} = \sum_i^{S_B} k'_{ij}. \tag{C.19}$$

The procedure produces a new contingency table that has the same number of entries as the original one. For each distribution $G_A(s)$ that we considered, we used the procedures described in section C.1.1 to estimate the mutual information for that particular distribution. The maximum mutual information over all signal distributions calculated is our estimate of the channel capacity C. As mentioned above, the errors reported for C represent the 95% confidence interval for the intercept estimated by the linear extrapolation procedure (section C.1.1.2).

C.2 Additional Experimental Calculations

C.2.1 Control calculations

As mentioned in the main text, we examined the channel capacities between the activities of the initiator caspase (IC; cleaved caspase 3) and both the effector caspase (EC; cleaved PARP) and

terminal cellular phenotype. We found that the IC to EC channel capacity exceeded 1.2 bits. This confirms the nature of IC as an intermediate component in the TRAIL signaling network due to the relative increase in information when using IC as the input distribution to the channel capacity calculation instead of TRAIL.

C.2.2 Population size dependence of single-cell channel capacity

Given our large data set, we investigated how channel capacity would vary for individual cells as a function of population size; as mentioned in the main text, the population-level channel capacity has a clear dependence on the size of the population. We expected to find that as the sample size increases the estimators describing the response distribution will be sufficiently accurate to prevent the need to calculate the channel capacity from the entire data set of over 1.2 million cells (which is computationally expensive). We confirmed this empirically, upon calculation of the single-cell channel capacity for increasing subsets of our FACS-generated data, using sample sizes of 500, 1000, and 2000 cells per TRAIL concentration.

C.2.3 Dose-dependent scaling

In our data we observed that IC activity levels were substantially higher in dead than live cells, most likely due to the variety of positive feedback mechanisms present in the caspase cascade [155]. Because this additional cleavage of caspase 3 in dead cells occurs downstream of the cell's commitment to apoptosis, it could be considered a consequence of the cell's phenotypic outcome rather than as an intermediate factor contributing to it. Since the channel capacity estimation is time-dependent (and capturing the exact moment of cell death for every cell is technologically infeasible), we proceeded to examine the impact that post-commitment IC activity has on our estimates for channel capacity between TRAIL dose and IC activity level. We therefore performed our analysis separately for live and dead cells by partitioning them into these two groups according to the threshold effector caspase response, t_{EC} [99]; we calculated this quantity by estimating the minimum density between the two peaks of the bimodal EC activity distribution: $\log_{10}(t_{EC}) =$

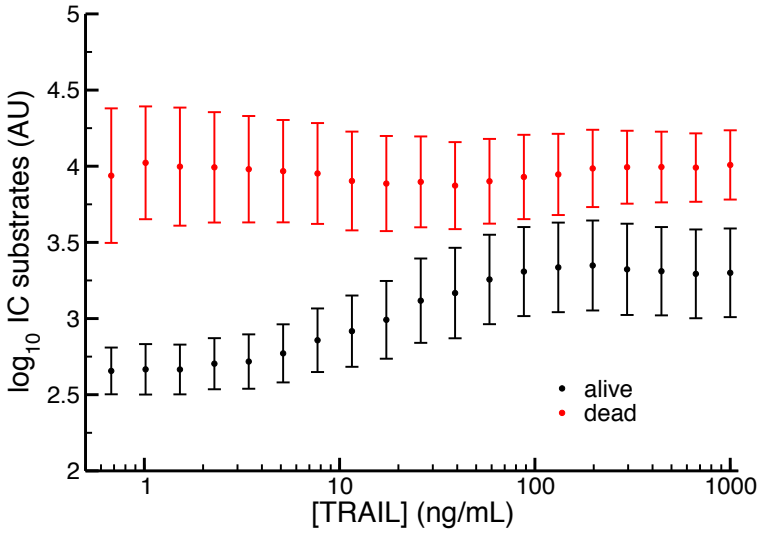


Figure C.3: Initiator caspase activity scales with TRAIL among living cells. Shown for each TRAIL concentration are the sample mean and standard deviation ($n \approx 60,000$ cells)

2.85 ± 0.05 [113]. We then plotted the dose response data to determine how response varies with TRAIL concentration. These plots show clearly that only initiator caspase activity scales with TRAIL dose, and it does so only among living cells (Figure C.3). The channel capacity between TRAIL and IC activity in living cells is approximately 1.01 bits as shown in Table 1 in the main text, essntially the same as the channel capacity between TRAIL and IC activity in all cells. Mean effector caspase activity in living and dead cells in addition to mean initiator caspase activity in dead cells does not significantly vary for differing doses of TRAIL (Figures C.3 and C.4) and as a result, we did not calculate the channel capacity for these dose-response relationships.

C.2.4 Resampling experimental data

In order to calculate the channel capacity for the pheromone signaling network in yeast, we reconstructed dose-response data shown by Bashor *et al.* in their Figure 4D [78]. In this case the signal distribution was a set of logarithmically spaced α-factor concentrations and the corresponding response was *pFUS1*-GFP fluorescence. With this data we constructed a series of dose-dependent

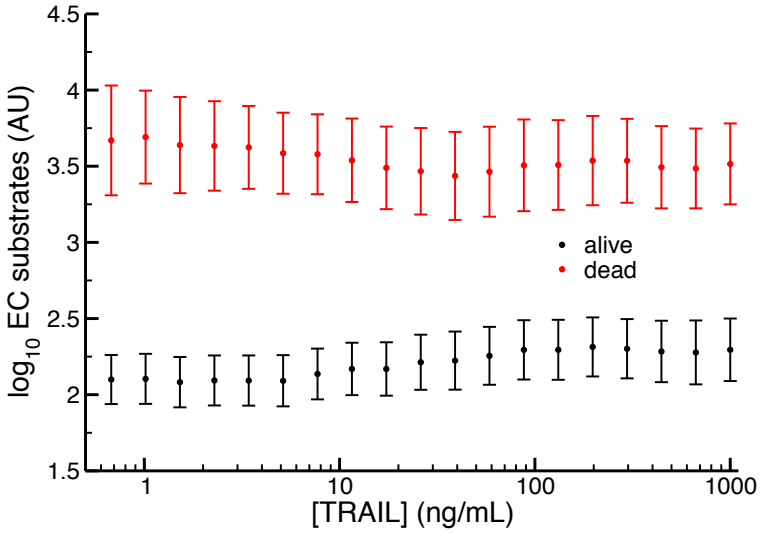


Figure C.4: Effector caspase activity is invariant with TRAIL among both living and dead cells. The data set is identical to that in Figure C.3.

Gaussian distributions defined by the mean and standard deviation of the *pFUS1*-GFP response given some α-factor concentration. From these distributions we sampled 100 values for each of 10 α-factor concentrations in order to construct a dose-response data set from which we could estimate the channel capacity (Figure C.5). We similarly performed this procedure for calculating the population-level channel capacity for the set of MCF10A and HeLa cells shown in Figure 3B of the main text. In this case, the mean and standard deviation for a particular TRAIL dose refer to the number of living cells in a given population.

C.3 Spatial Channel Capacity

In order to calculate the spatial channel capacity between a motile cell undergoing chemotaxis and its target that is producing some chemical gradient, we constructed signal/response pairs from angles between the cell and its target. We used the CellTrack program developed by Sacan *et al.* [121] to output text files containing frame-by-frame coordinates for the edges of the cells and their centers of mass (COM). We used these coordinates to calculate time-dependent signal and

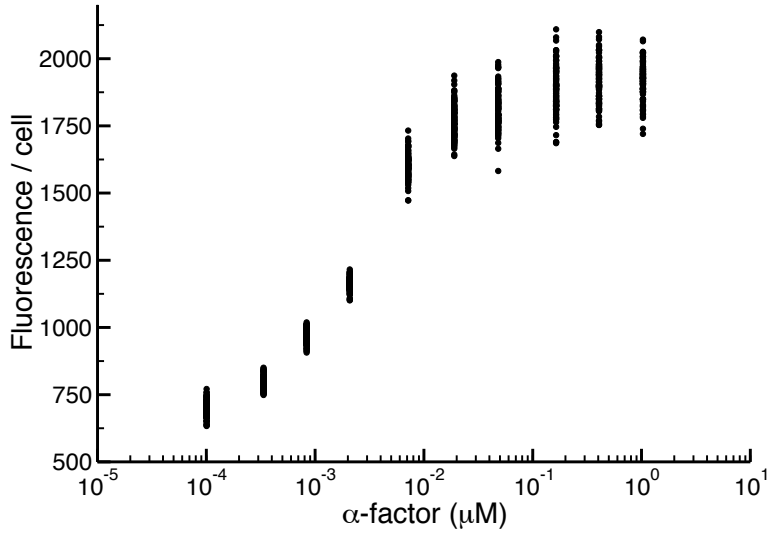


Figure C.5: Resampled data from Figure 4D in Bashor *et al.* [78].

response angles.

C.3.1 Neutrophil motion

We initially analyzed the motion of a neutrophil that is “chasing" a bacterium from a classic movie taken in the 1950s (see supplemental files, or this website). For the purposes of our calculation, we assume that the neutrophil is in fact following a chemical gradient generated by the bacterium. In this case, the signal corresponds to the angle, termed θ_1, between the bacterium at a particular frame in the movie, time t, and the neutrophil at another time $t + \Delta t_1$. The subsequent response angle, θ_2, is that of neutrophil motion between time $t + \Delta t_1$ and time $t + \Delta t_1 + \Delta t_2$. These angles then comprise the signal and response distributions used to calculate the channel capacity of the system. A visual representation of this calculation can be seen in the main text (Figure 4).

We calculated the signal and response angles between the neutrophil COM and bacterial COM relative to the x-axis unit vector in the Cartesian coordinate system. This method is similar to one outlined by Burov *et al.* derived to provide more directional information than mean squared displacement for analysis of random walks [122]. We employ a “windowed" data collection method;

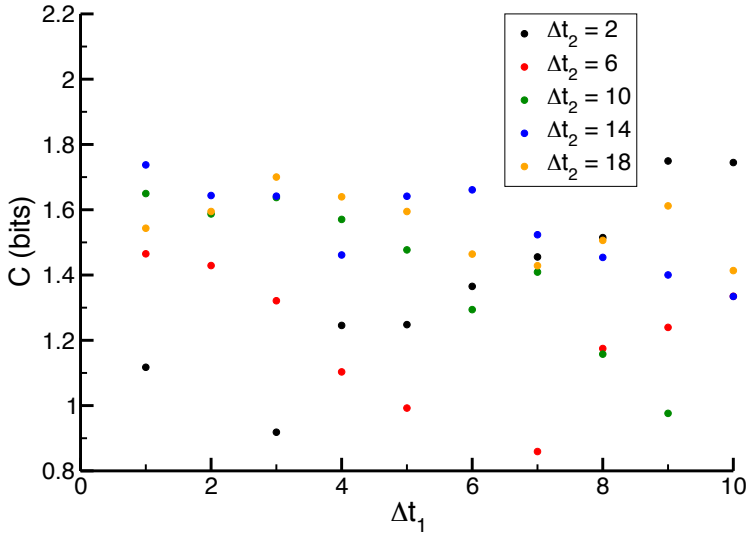


Figure C.6: Channel capacity as it depends on Δt_1 and Δt_2 for select values.

given some starting time, t, we calculate an arbitrary signal and response angle pair, requiring information from time points, $t+\Delta t_1$ and $t+\Delta t_1+\Delta t_2$. In our windowed data collection, the next pair of angles is calculated using t incremented by one frame: $t = t+1$. To confirm that the calculated channel capacity was not an artifact of the chosen time delay values, Δt_1 and Δt_2, we explored the nearby $(\Delta t_1, \Delta t_2)$-space and discovered that the channel capacity is relatively robust to Δt_1 and Δt_2 as seen in Figure C.6.

C.3.2 *Dictyostelium* motion

The next movie we analyzed is that of a *Dictyostelium* cell following a cAMP gradient (see supplemental files or this website). In this movie, *Dictyostelium* responds to cAMP introduced by a pipette tip which changes location periodically. Since the pipette tip remains stationary between location shifts, we can employ our original calculation used for the neutrophil/bacterium data and omit the Δt_1 parameter (the largest channel capacity occurs when $\Delta t_2 = 18$). This omission is valid since θ_1 is identical for a range of θ_2 values (*i.e.* no motion in the gradient source/change in the signal).

C.4 Simple Model

As mentioned in the main text, the initial model takes the form:

$$R = (R_{max} - R_{min}) \cdot \frac{S^n}{S^n + K^n} + R_{min} + \varepsilon \tag{C.20}$$

where the normally-distributed noise term $\varepsilon \sim N(0, \sigma)$ depends on some chosen standard deviation, σ. The parameter values chosen for the base model (shown in Figure 3) are as described in the Materials and Methods section of the main text: $K = 10$, $n = 6$, $R_{max} = 30$, and $R_{min} = 20$. For all models discussed in the paper, the response threshold governing an individual cell's fate is positioned such that half of the signal values produce mean responses below the threshold and half produce mean responses above the threshold.

C.4.1 Choosing signal values

We selected evenly-spaced signal values to achieve responses 10% above the minimum response and 10% below the maximum response (the *transition zone*):

$$0.1 \cdot (R_{max} - R_{min}) + R_{min} \leq R \leq 0.9 \cdot (R_{max} - R_{min}) + R_{min} \tag{C.21}$$

Through simple algebra we show that the resulting minimum and maximum signal values, (S_{min} and S_{max}, respectively) are:

$$S_{min} = K \cdot \sqrt[n]{\frac{1}{9}} \tag{C.22}$$

and

$$S_{max} = K \cdot \sqrt[n]{9} \tag{C.23}$$

This prevents selection of signal values that would produce extremely high or low responses, since sampling more of these responses relative to intermediate responses would reduce the channel capacity.

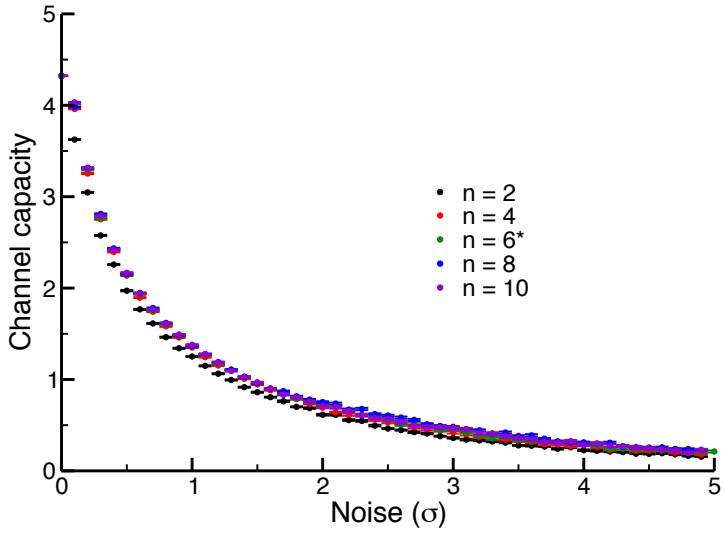


Figure C.7: Single-cell channel capacity with respect to noise for a range of n values. The starred $n = 6$ denotes the value used in all other calculations based on this model. There is minimal difference between models where $n > 2$ and even the model with $n = 2$ displays qualitatively similar behavior to the others. Error bars denote 95% confidence about the intercept estimate (see section C.1.1.2)

C.4.2 Varying n

In order to determine the effect of our chosen $n = 6$ on this model's channel capacity (both single-cell and population-level), we varied n between 2 and 10. We see in general from Figures C.7 and C.8 that this variation produces minimal difference between models; qualitatively, models with different n are nearly identical.

C.4.3 Channel capacity saturation with population size

As discussed in the main text, the fraction of a group of cells making a particular signal-dependent decision is the statistic used to determine collective response for the population-level calculation of mutual information. By calculating this statistic over a number of replications, we effectively construct a sampling distribution for the fractional response to some arbitrary signal. Since the standard deviation of this distribution (*i.e.* the standard error of the statistic) is dependent on sam-

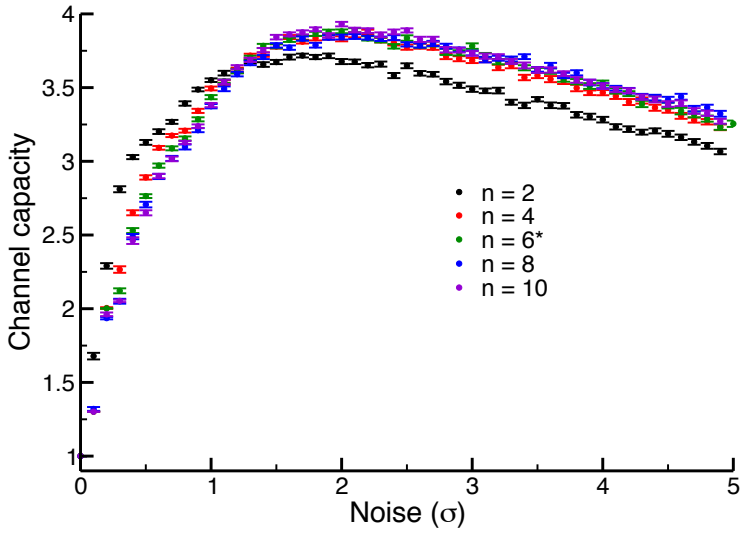


Figure C.8: Population-level channel capacity with respect to noise for a range of n values. Again we see little difference between models with different n with the minor exception of $n = 2$. We do observe a slight shift in the amount of noise producing maximal channel capacity, but the qualitative trends are essentially identical. Error bars are as in Figure C.7.

ple size, we observe an inverse correlation between the size of the population and the standard error of the fractional response (Figure C.9). If we restrict our data set to those values in the increasing regime of the dose-response curve (the *transition zone*, see Section C.4.1), increasing the population size results in the channel capacity approaching its theoretical maximum of $C = -\log_2(\frac{1}{N})$ bits (*i.e.* the entropy of the signal distribution in the transition zone) where N is the number of signal values in the transition zone. We observe this channel capacity saturation in Figure 3D of the main text.

C.4.4 Maximal fractional response

At high levels of noise, we observe another interesting feature of the population response: the inability to effect a universal population response at arbitrarily high signal levels. In other words, no matter how much signal is present in the environment, there will still be a fraction of cells in a population that does not respond. This is plainly observed graphically in Figure C.10; even as the

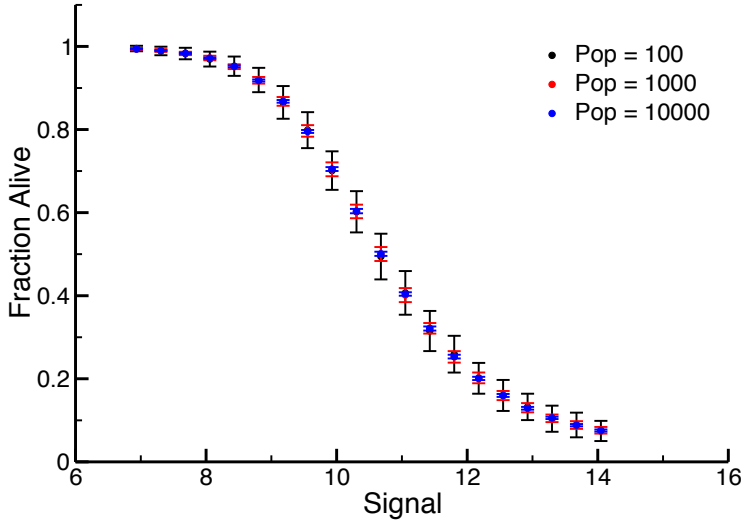


Figure C.9: Population-level dose-response curves for multiple population sizes

response saturates at low and high signal levels, there is sufficient noise such that a subpopulation of cells at these signal levels fall above and below the threshold, respectively. The corresponding population dose-response curve thus exhibits saturating, incomplete responses both at low and high signal levels (Figure C.11).

C.4.5 Population channel capacity dependence on signal spacing

As mentioned previously, we restricted sampling signal space (on the individual cell level) to the region generating responses between 10% and 90% of the model's maximum response, since the majority of the information resides in this section of the single-cell dose-response curve (see Section C.4.1). However, as the noise decreases on the single-cell level, the shape of the population dose-response curve changes, becoming more switch-like and ultimately shrinking the signal range across which the population-level transition occurs (Figure C.12). If we then engage in a similar strategy for the population dose-response curve by sampling a fixed number of signal values corresponding to responses between 10% and 90% of the maximal *population* response for a given noise value, we can keep the population-level channel capacity constant as noise approaches 0,

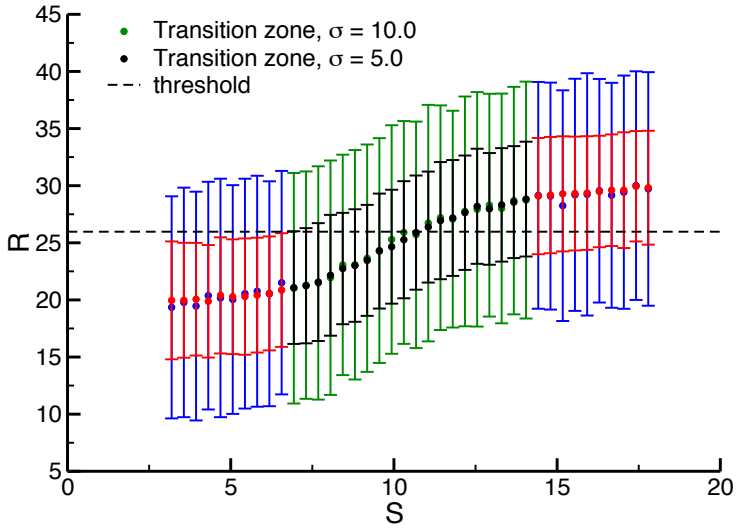


Figure C.10: Single cell response curve at high noise values

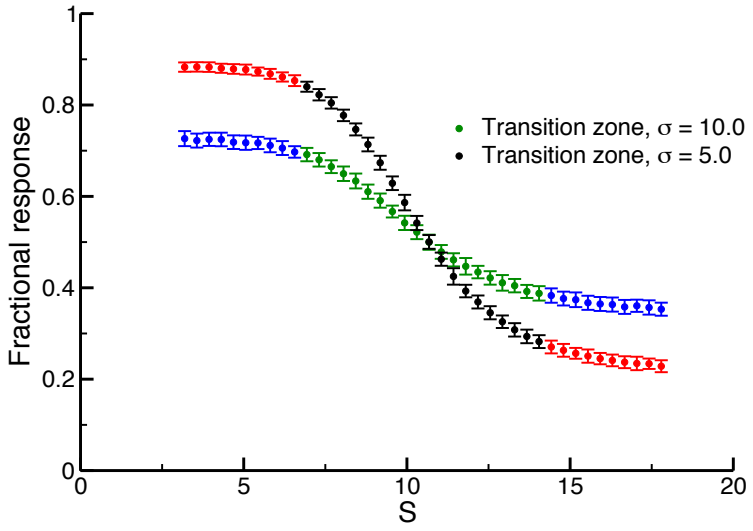


Figure C.11: Population response curve at high noise values (population size = 1000)

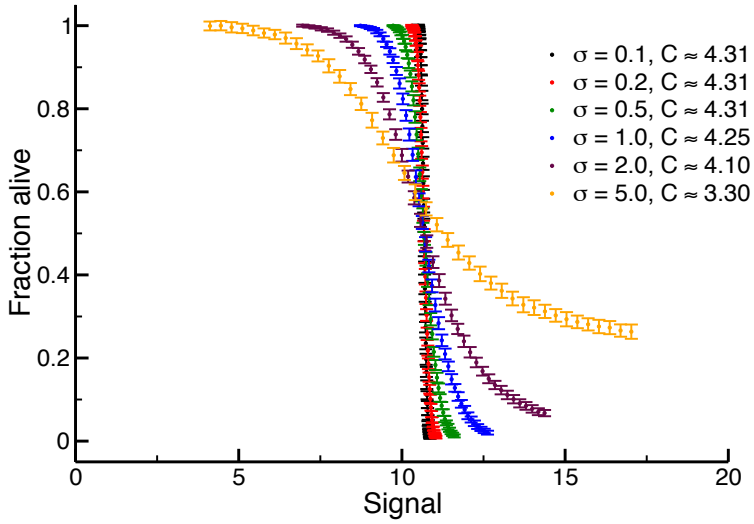


Figure C.12: By sampling evenly-spaced signal values in the population dose-response transition zone (which varies given some level of noise in the individual cell), we observe that the population-level channel capacity can be maintained at a constant value as noise approaches 0.

as shown in Figure C.12. It thus appears that generating high population-level channel capacity requires only an exceedingly small level of noise, so long as the number of signal values sampled remains constant across the (potentially very narrow) region over which the majority of the transition occurs. A similar numerical experiment reveals that by simply increasing the number of signal values sampled in the original signal range (defined by the signal values corresponding to the single-cell transition region), the amount of noise required to reach the maximal population-level channel capacity decreases (Figure C.13). It is thus unclear why cells might have evolved high levels of noise to control population-level responses, when these results suggest that any non-zero level of heterogeneity would suffice.

It is important to note, however, that maintaining a high population-level channel capacity for arbitrarily low levels of noise requires increasingly smaller spacing between individual signal values (see Figure C.12). This is problematic since the existence of variability in the signal itself interferes with very small signal spacing in physically realistic systems. For example, if we consider an *in vivo* scenario in which hormones or cytokines are distributed to the various tissues of an

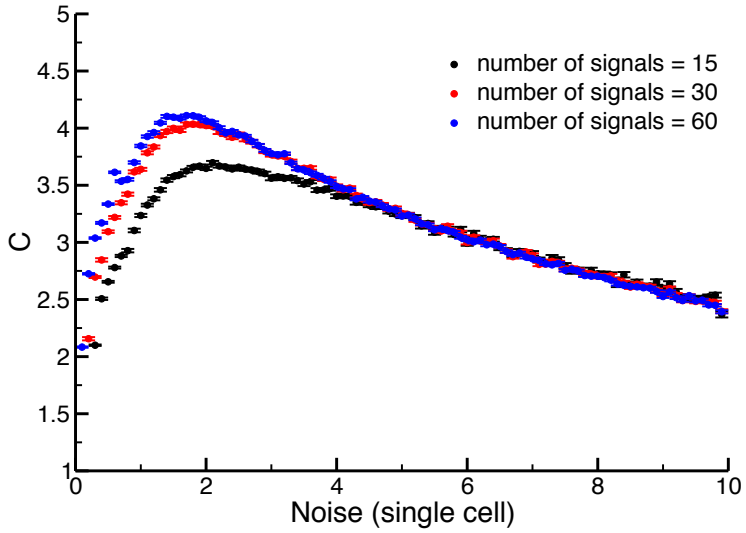


Figure C.13: Increasing the number of sampled signal values in the transition zone of the *individual* cell's dose-response curve (*i.e.* independently of the population dose-response transition zone) results in a decrease in optimal noise level and an increase in channel capacity which appears to approach some limit.

organism, we know that these molecules are themselves produced by other cells. Since cytokine production is a stochastic process, the amount of signal to which specific cells within a tissue are exposed will be a variable quantity.

To confirm that the existence of variability in the signal does in fact produce optimal population-level channel capacity at non-trivial levels of single-cell noise, we introduced another noise term (the *signal* noise, ε_s, as opposed to the original *response* noise, ε) governing the limit of signal accuracy for our populations of simulated cells. As an example, consider signal noise equal to 1% of the signal value that produces an average response corresponding to the decision-making threshold. This alternate form of the model (modified from Equation C.20) has the following form:

$$R = (R_{max} - R_{min}) \cdot \frac{(S+\varepsilon_s)^n}{(S+\varepsilon_s)^n + K^n} + R_{min} + \varepsilon \tag{C.24}$$

where the signal noise term ε_s is normally distributed and is sampled independently for each *population* of cells: $\varepsilon_s \sim N(0,\sigma)$. This procedure simulates the previously discussed example of

cytokine production and distribution to cellular populations, and it is distinct from the application of the original noise term, ε, that was applied to each individual cell. We then calculated the population-level channel capacity for data sets in which the *population-level* transition zone is fixed, by altering the signal space density so that the responses corresponding to these signals fall between 10% and 90% of the maximal population response. Using this data, we characterized the impact of signal detection limits on the population-level channel capacity with respect to different levels of ε (Figure C.14). As expected, we see that slight error in the signals to which the populations are exposed reduces the population-level channel capacity at low levels of response noise. Thus, although very small levels of single cell noise could theoretically produce high population-level channel capacity, the presence of signal variability in physically realistic systems requires higher levels of noise, so that signal values can be spaced at reasonable distances from one another while maintaining low variability in the population-level response.

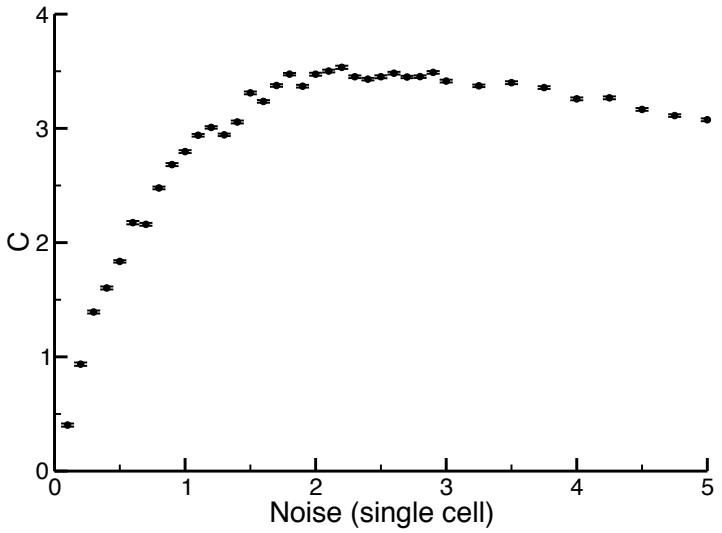


Figure C.14: Implementing a limit on signal resolution in the form of a fixed level of signal noise (ε_s) results in a positive correlation between noise level and channel capacity on the population level. In this figure ε_s is one percent of the signal value corresponding to the decision-making response threshold. We also note that this data exhibits a population-level $C < 1$ bit when the single-cell noise is 0, and that this differs from the data in the main text and Figure C.13. This results from the presence of relatively high signal noise given the sharp transition region that occurs in the population response with negligible single cell (response) noise (Figure C.12) as compared to the lack of signal noise in other mentioned data sets.

Appendix D

Appendix for Chapter 4

D.1 Framework

D.1.1 Model with low Hill coefficient

Shown in Figure D.1 are the channel capacity as a function of the signal window (A) and the number of sampled signal values (B). These only differ from those in the main text in terms of the Hill coefficient used to generate the data. These plots have Hill coefficients of 6 as opposed to those in the main text that had Hill coefficients of 60. As can be clearly seen, there is no significant difference between the two data sets.

D.1.2 Varying the transition zone bounds

We performed a brief analysis of the simple model to characterize the effects of varying the response range of the transition zone that is shown in Figure D.2. In the main text all transition zones are constructed using bounds of 10% and 90% maximal response (after subtracting basal response). We thus varied the percentage of response space removed from consideration to examine its effect on the transmission of information in this simple model. Though there is some variation, the trends are relatively flat and appear to depend somewhat on the variability in response. As a result, we will maintain our focus on the 10%-90% transition zone with the caveat that modifying

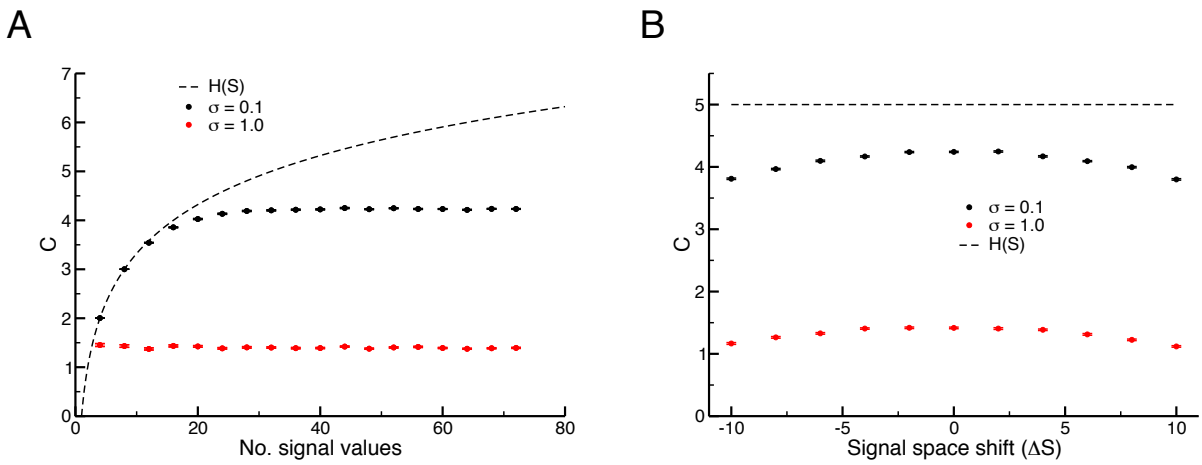


Figure D.1: A & B. Similar to Figure 5.1C & D in the main text, but the data was generated with a Hill coefficient of 6 instead of 60.

this parameter could introduce minor fluctuations in the resulting channel capacity values.

D.1.3 Finding the transition zone empirically

For the majority of stochastically simulated models, we empirically determined the bounds of the transition zone by sampling a range of values that span the increasing regime of the dose-response curve and appear to approach the minimum and maximum responses. In the case of the kinase cascade models and the covalent modification cycles, the dose-response curves were sufficiently sigmoid in nature to allow a least squares regression fit to a Hill function. The resulting functional fit was inverted to find the signal values which bound the transition zone. Since the recruitment of Sos to the EGFR molecule is not so well represented by a simple sigmoid function, we took a different, but equally straightforward approach. We similarly characterized the majority of the dose-response curve in signal space and proceeded to use linear interpolation to estimate values corresponding to the signal bounds of the transition zone.

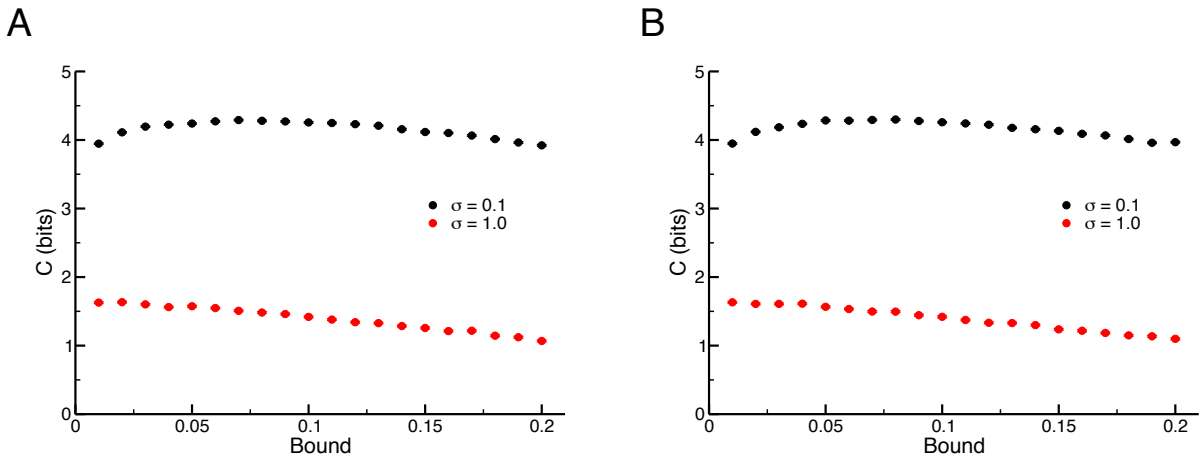


Figure D.2: Varying the relative width of the transition zone shows minimal impact on the simple model regardless of utilizing high (A) or low (B) Hill coefficients in generating the data. The x-axis shows the fraction of response values that are removed from both high and low ends of response space (*i.e.* if the response interval is [0,10] and the bound is set to 0.1, then the transition zone interval is [1,9]). As can be seen, the choice of bound impacts the information estimation to some degree, but not sufficiently to warrant the computationally intensive search for the bounds optimizing each calculation. We use a bound of 0.1 or 10% throughout this work, and in this example the deviation from the optimal values is marginal at best.

D.2 Binary interaction model

Figure D.3 shows how the number of sampled signal values in the transition zone alters the estimated information transmission in the LT model that includes synthesis and degradation of both ligand and receptor components.

D.2.1 Analytical solution for transition zone with molecular turnover

The system of ordinary differential equations for the LT model is:

$$\frac{\mathrm{d}L}{dt} = k_-[B] - k_+[L][T] - \delta_L([L]+[B]) + Q$$
$$\frac{\mathrm{d}T}{dt} = k_-[B] - k_+[L][T] - \delta_T([T]+[B]) + Q$$
$$\frac{\mathrm{d}B}{dt} = -k_-[B] + k_+[L][T] - (\delta_L+\delta_T)[B]$$

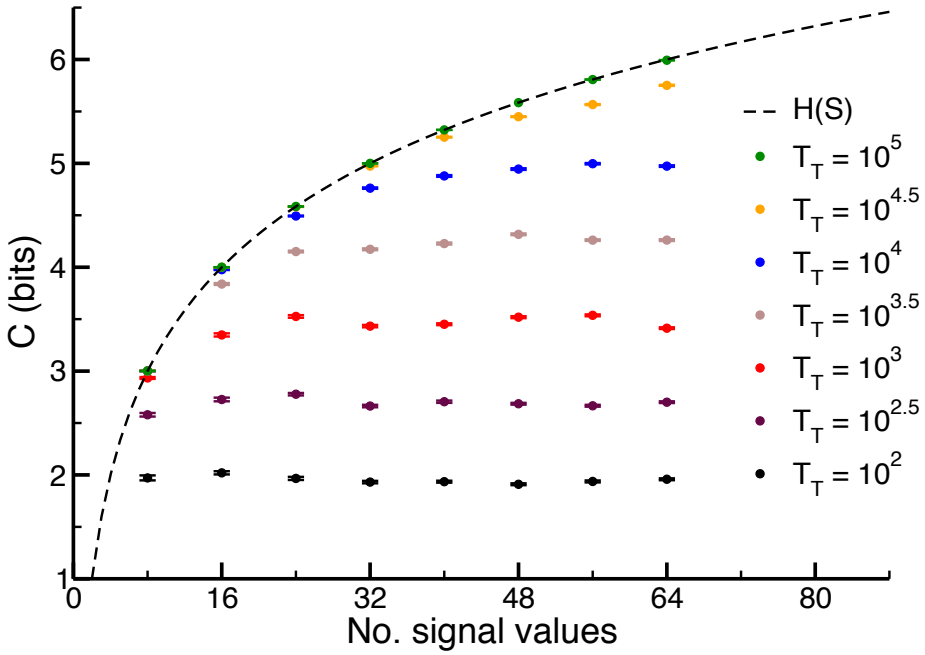


Figure D.3: Similar to Figure 5.2C in the main text, but using the LT model that includes rules for synthesis and degradation.

where L,T, and B are the ligand, receptor and bound complex concentrations, δ_M and Q are the degradation and synthesis rates of some molecule M, and k_- and k_+ are unbinding and binding rates. Note that total ligand and receptor concentrations are defined as $[L_T] = [L] + [B]$ and $[T_T] = [T] + [B]$ and that synthesis of L and T are equivalent, but not degradation . Furthermore, we parameterized our model such that $\frac{Q}{\delta_M} = M_T$. We can then solve for B as a function of L_T at equilibrium and invert the equation to determine the transition zone bounds, where $B_{\max} = 0.9 \cdot T_T$ and $B_{\min} = 0.1 \cdot T_T$ are the maximum and minimum responses in the transition zone, respectively. The equation for B given L_T is quadratic:

$$0 = k_+[L_T][T_T] - (k_+[L_T] + k_+[T_T] + k_- + \delta_L + \delta_T)[B] + k_+[B]^2 \tag{D.1}$$

and finding L_T is trivial, with knowledge of the parameter values.

D.3 Covalent modification cycle model

D.3.1 Varying signal

In the seminal characterization of the covalent modification cycle, Goldbeter & Koshland employed a ratio, here termed r, to serve as the input to the system [87]. This ratio is defined as the maximum velocity of the activating enzyme (kinase) over the maximum velocity of the deactivating enzyme (phosphatase):

$$r = \frac{k_{cat,K} \cdot [K_T]}{k_{cat,P} \cdot [P_T]} \tag{D.2}$$

where $[K_T]$ and $[P_T]$ is the total concentration of kinase and phosphatase, respectively. This quantity is then varied logarithmically to produce dose-response trends. Traditionally, r is varied by modifying the copy number/concentration of the kinase, however we observed that for certain parameter regimes, $S_{\min}$ was sufficiently low, such that a logarithmic distribution of arbitrary numbers of signal values could not be realized in integer space. Therefore, we varied r by modifying the catalytic rate of the phosphatase: $k_{cat,P}$. To fix the level of saturation of the phosphatase (which

is defined by the Michaelis constant, $K_{M,P}$) to be equal to the level of saturation of the kinase, we co-varied the association rate of the active substrate and phosphatase:

$$k_{on,P} = \frac{k_{off,P} + k_{cat,P}}{K_{M,K}}. \tag{D.3}$$

D.4 Kinase cascade models

Shown in Figure D.4 are the raw dose-response data sets for the solution and scaffold models. It is important to note that the results observed, specifically the very low activation of the scaffold model, is a result of the parameterization. Since the two models' signals are given by the copy number of an initiator agent, the lowest possible signal is $S = 1$. Due to the sensitivity of high-depth solution models to signal, we parameterized the models to suppress signal throughput by increasing the copy number of the phosphatase corresponding to K_1 and assigned a low association rate to all binding events. While this produced an appropriate transition zone for a solution model with a depth of 4, it also suppressed output in the scaffold model, resulting in increased noise due to stochastic effects. Thus, the important result to take away from these models is not that the solution model exhibits higher information, but the overall trends of the information transmission within each model (*e.g.* dose-response alignment preserves information transmission for network intermediates).

Seen in Figure D.5 is the channel capacity between signal and the activity of the final kinase in the cascade as a function of the total depth of the cascade. We clearly observed that the solution model exhibits consistently higher information transmission in this scenario and that both signaling paradigms exhibit degradation of information as the depth of the cascade increases.

D.4.1 Model parameters

In order to facilitate comparison between the scaffold and solution models, we chose parameter values, such that models of arbitrary depth (*i.e.* distinct kinase types) would have consistent pa-

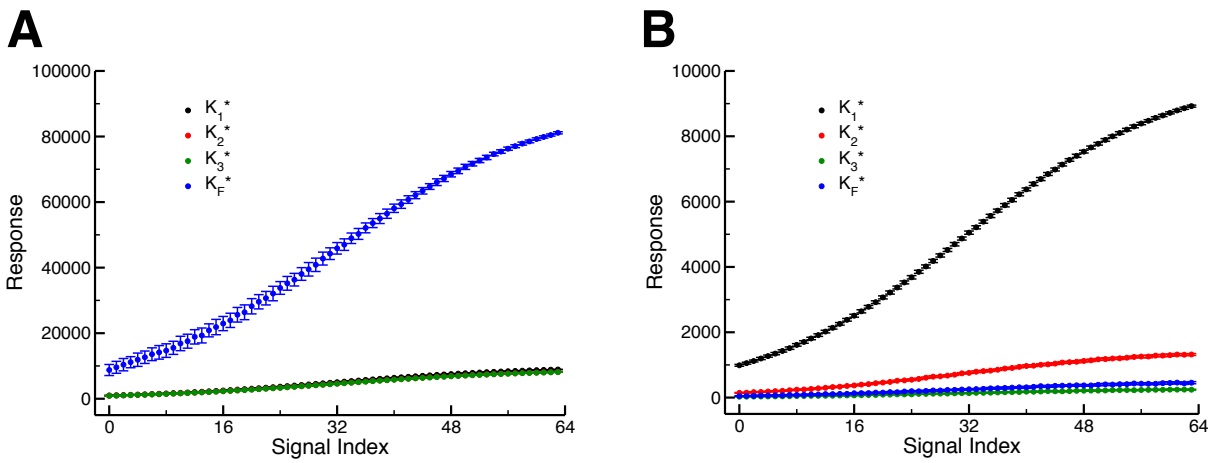


Figure D.4: (A) Dose-response data for the solution model using the VTZ approach. Though difficult to see, all observables approach 90% activation. (B) As (A) but for the scaffold model. The trends here are distinct from (A), where only K_1 approaches 90% activation. The other observables, including K_F, fall well short of maximal activation.

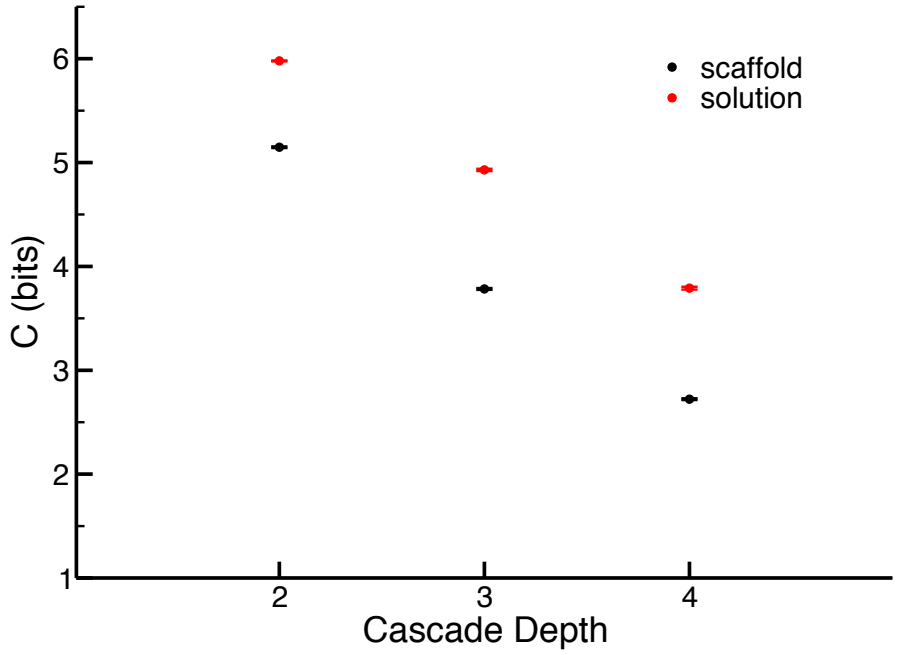


Figure D.5: Information transmission to final kinase ($C(S;K_F)$) for various cascade depths

Parameter	Value
Association	$10^{-7}\,(\text{molec}\cdot\text{sec})^{-1}$
Dissociation	$0.1\ \text{sec}^{-1}$
Catalysis	$1\ \text{sec}^{-1}$
Kinase $(1 \leq i < d)$	10^4 copies
Kinase (d)	10^5 copies
Scaffold	10^4 copies
Phosphatase (1)	10^5 copies
Phosphatase $(1 < i \leq d)$	10^3 copies

Table D.1: Parameter values for the scaffold and solution models, where d denotes the depth of the cascade.

rameter values. Since the solution model exhibited relatively high sensitivity to signal (as seen in the main text), the association rates for kinase-kinase interactions had to be relatively slow, though still in a biologically relevant regime. Table D.1 shows the values used in the models.